Extended Abstracts

presented at the

25th International Symposium on Mathematical Theory of Networks and Systems MTNS 2022

held

12-16 September 2022

in

Bayreuth, Germany,

and

edited by

M. H. Baumann

L. Grüne

B. Jacob

K. Worthmann

MTNS 2022 featured full paper and extended abstract submissions. This part of the proceedings contains the extended abstracts. The full papers presented at MTNS 2022 are available online at https://www.sciencedirect.com/journal/ifac-papersonline/vol/55/issue/30 as Volume 55, No. 30 of IFAC PapersOnline.

The 25th International Symposium on Mathematical Theory of Networks and Systems MTNS 2022 was **co-sponsored** by:

- International Federation of Automatic Control (IFAC)

- Deutsche Forschungsgemeinschaft (DFG; German Research Foundation)

- Oberfrankenstiftung

- University of Bayreuth

Additionally, for their generous support, **thanks** go to:

- Bechtle AG
- Foundation Advancement of Mathematics

The complete lists of program committee members, organizers, etc. can be found in the IFAC PapersOnline volume of MTNS 2022 https://doi.org/10.1016/S2405-8963(22)02792-6.

Foreword

After more than two years of limited social and scientific interactions due to the Covid-19 pandemic, it was a pleasure to welcome more than 300 participants in person and about 60 online participants at MTNS 2022 in Bayreuth. Submissions to MTNS 2022 were possible as extended abstracts and full papers. The accepted full papers that were presented at the conference are published in IFAC PapersOnline https://www.sciencedirect.com/journal/ifac-papersonline/vol/55/issue/30. In this volume you find the extended abstracts that were presented at the conference. Further, you also find the titles of the plenary and semi-plenary talks as well as their abstracts resp. links to the corresponding full papers.

We hope you enjoy these abstracts and to see you in person at MTNS in the future.

The Editors
M. H. Baumann, L. Grüne, B. Jacob, and K. Worthmann

Plenary Talks

- Claudio De Persis (University of Groningen, The Netherlands)

On data-driven control

Abstract. We present a technique to design controllers from data for systems whose model is imprecisely known. The technique is based on collecting measurements of low complexity from the systems and using them for the synthesis of controllers, which is reduced to the solution of data-dependent semidefinite programs. The method provides stability certificates in the presence of perturbations on the dataset.

- Luz de Teresa (National Autonomous University of Mexico, Mexico City, Mexico)

Some results on hierarchical control for parabolic equations

Abstract. In classical control theory, we usually have a state equation or system and just one control, with the mission of achieving a predetermined goal. Sometimes, the goal is to minimize a cost function in a prescribed family of admissible controls; this is the optimal control viewpoint. A more interesting situation arises when several (in general, conflictive or contradictory) objectives are considered. This may happen, for example, if the cost function is the sum of several terms and it is not clear how to average. It can also be expectable to have more than one control acting on the equation.

In this talk, we present an overview of the known results on this subject for the heat equation. We will recall the results of Araruna and collaborators where hierarchic exact controllability results were established for linear and semilinear heat equations. In this research, and in the seminal papers by J.-L. Lions, the main idea is to work with one primary control (the leader) and one or several secondary controls (the followers). For each possible leader, the associated followers try to minimize a functional (or reach equilibrium if there is more than one cost objective function). Then, the leader is chosen such that the associated state satisfies a final time constraint. We will present the recent result with E. Fernández-Cara et al., where we accomplish optimal control and controllability tasks with a hierarchy of controls. This time, however, the controllability goal will be commended to the follower, while the choice of the leader will be subject to an optimal control problem. It will be seen that this makes the problem more difficult to handle (essentially because we must work all the time in a very restrictive class of leader controls).

- Weinan E (Peking University, China and Princeton University, NJ, USA)

Deep Learning and Optimal Control

Abstract. There is a close analogy between deep learning and optimal control. This analogy can be exploited to develop deep learning-based algorithms for optimal control, and optimal control-based algorithms for deep learning. I will discuss the progress made along these directions.

Full paper *Weinan E, Jiequn Han, Jihao Long* "Empowering Optimal Control with Machine Learning: A Perspective from Model Predictive Control" available in the IFAC PapersOnline volume of MTNS 2022 https://doi.org/10.1016/j.ifacol.2022.11.039.

- Maria Elena Valcher (University of Padova, Italy)

Opinion dynamics models: A mathematical abstraction of individuals' behaviours and interactions

Abstract. The talk will focus on the main mechanisms influencing opinion dynamics, like homophily, mutual appraisal, and bounded confidence. Some classic opinion dynamics models, as well as some recent ones, will be presented. Interesting open problems as well as promising research directions will be proposed.

- George Weiss (Tel Aviv University, Israel)

Lax-Phillips semigroups for nonlinear systems

Abstract. We briefly recall the basics about Lax-Phillips semigroups for well-posed linear systems, and the definition of well-posed nonlinear systems via nonlinear Lax Phillips semigroups. Then we concentrate on two results concerning well-posed nonlinear systems:

We investigate a special class of nonlinear systems that are obtained by modifying the second order differential equation that is part of the description of conservative linear systems "out of thin air" introduced by M. Tucsnak and G. Weiss in 2003. The differential equation contains a nonlinear damping term that is maximal monotone and possibly set-valued. We show that this new class of nonlinear systems is incrementally scattering passive (hence well-posed). Our approach uses the theory of maximal monotone operators and the Crandall-Pazy theorem about nonlinear contraction semigroups, which we apply to the Lax-Phillips semigroup of the system.

We investigate the class of incrementally scattering passive nonlinear systems, as defined in some earlier papers of ours. We show that these can be defined by a differential inclusion and a formula defining the current output in term of the current state and the current input. Our approach uses the theory of maximal monotone operators.

The talk is based on joint work with Shantanu Singh.

Semi-Plenary Talks

- Roland Herzog (University of Heidelberg, Germany)

The role of the metric in numerical linear algebra and optimization

Abstract. Many algorithms in everyday use implicitly employ the Euclidean inner product of the underlying space. While this is convenient and user-friendly on the one hand, it also turns out that the Euclidean metric may not be the one yielding the best performance of the respective algorithm. In this talk we revisit the role of the metric in a number of well-known algorithms in numerical linear algebra and optimization, and demonstrate the potential of user-defined metrics in each case.

- Anna-Lena Horlemann-Trautmann (University of St. Gallen, Switzerland)

The densities of good codes in various metric spaces

Abstract. The densities of codes with certain properties have always been of interest in classical coding theory, in particular to understand how many of such codes exist and how likely a random code will have the prescribed properties. Further applications of density results of codes appear in code-based cryptography, where it is important that the set of codes with a certain property is large enough to outgo brute force attacks. In this talk we will present various density results for optimal or close-to-optimal codes in different metric spaces with different types of linearity. In particular, we will show when optimal codes in the Hamming, rank and sum-rank metric are dense and when they are sparse.

- Boris Houska (ShanghaiTech University, China)

Global Optimal Control: Opportunities and Challenges

Abstract. Optimal control theory, algorithms, and software for analyzing and computing local solutions of linear and nonlinear optimal control problems have reached a high level of maturity, finding their way into industry. In the context of many applications, locally optimal control inputs can be computed within the milli- and microsecond range. This is in sharp contrast to the development of algorithms for locating global minimizers of non-convex optimal control problems, which is hindered by several key issues, including the overall complexity of generic optimal control problems and their curse of dimensionality. This talk reviews and discusses recent solutions that address these rather fundamental challenges including novel types of Branch & Lift methods as well as modern Koopman-Pontryagin operator based lifting methods for global optimal control. Various numerical experiments will be used to illustrate the effectiveness of these approaches. The talk concludes with an assessment of the state of the art and highlights important avenues for future research.

- Achim Ilchmann (Technische Universität Ilmenau, Germany)

Funnel control – history and perspectives

Abstract. The control objective in funnel control is output feedback control such that the norm of the error e(t) of the closed-loop system remains inside a prespecified funnel with boundary $\varphi^{-1}(t)$, i.e. $\|e(t)\| < \varphi^{-1}(t)$ for all $t > 0$. In other words, prescribed transient behaviour as well as asymptotic accuracy is achieved. Typical features of funnel control are:

Simplicity of the feedback law. The feedback does not invoke any identification scheme, but is – for example in the relative degree one case – a time-varying error feedback of the form $u(t) = -1 / (1 - \varphi(t)\,\|e(t)\|)\, e(t)$, $e(t) := y(t) - y_{ref}(t)$, where $y_{ref}(\cdot)$ denotes a sufficiently smooth bounded signal with bounded derivative.
Note that the gain $k(t) = -1 / (1 - \varphi(t)\,\|e(t)\|)$ is large if, and only if, the error is close to the funnel boundary.

Funnel control is feasible for a whole class of input-output systems, which is characterized by structural assumptions, e.g., well-defined relative degree and stable zero dynamics.

After two decades of high-gain adaptive control, funnel control was introduced in 2002. First results were on linear, single-input, single-output, time-invariant systems with relative degree one and being minimum phase. From then on feasibility of funnel control was shown for other system classes such as multi-input, multi-output, nonlinear, infinite dimensional, perturbed systems, unknown control directions – provided they have stable zero dynamics and satisfy certain assumptions on the high-frequency gain. A particular challenge was to show feasibility for systems with higher relative degree, and to design a funnel controllers for systems described by partial differential equations. Funnel control was applied to various applications such as control in chemical reactor models, industrial servo-systems, wind turbine systems, electrical circuits, to name but a few. Recently, funnel control has been investigated in combination with model predictive control and applied to magnetic levitation systems.

- Christopher M. Kellett (Australian National University, Canberra, Australia)

Discontinuous Feedbacks for Stabilization and Combined Stabilization and Safety

Abstract. It has long been known that asymptotic controllability of a nonlinear system to a desired equilibrium or target set require discontinuous controllers for feedback stabilization, which, in turn, is equivalent to the existence of a nonsmooth control Lyapunov function. More recently, results combining stabilization and safety, captured by so-called barrier functions, have been proposed. This also gives rise to the need for discontinuous feedback controllers, though for slightly different reasons. In this talk, we summarise these results and present a hybrid feedback solution to the combined stabilization and safety problem for a non-trivial class of systems.

- Dante Kalise (Imperial College London, UK)

High-dimensional approximation of Hamilton-Jacobi-Bellman PDEs in deterministic optimal

Abstract. Optimal feedback synthesis for nonlinear dynamics -a fundamental problem in optimal control- is enabled by solving fully nonlinear Hamilton-Jacobi-Bellman type PDEs arising in dynamic programming. While our theoretical understanding of dynamic programming and HJB PDEs has seen a remarkable development over the last decades, the numerical approximation of HJB-based feedback laws has remained largely an open problem due to the curse of dimensionality. More precisely, the associated HJB PDE must be solved over the state space of the dynamics, which is extremely high-dimensional in applications such as distributed parameter systems or agent-based models.

In this talk we will review recent approaches regarding the effective numerical approximation of very high-dimensional HJB PDEs. We will explore modern scientific computing methods based on tensor decompositions of the value function of the control problem, and the construction of data-driven schemes in supervised and semi-supervised learning environments. We will highlight some novel research directions at the intersection of control theory, scientific computing, and statistical machine learning.

- Yann Le Gorrec (National Engineering Institute in Mechanics and Microtechnologies "ENSMM", Besançon, France)

Control design for distributed parameter systems – the port Hamiltonian approach

Abstract. This talk is concerned with the control of distributed parameter systems defined on a 1D spatial domain using the port Hamiltonian framework. We consider two different cases: when actuators and sensors are located within the spatial domain and when the actuator is situated at the boundary of the spatial domain, leading to a boundary control system (BCS). In the first case we show how dynamic extensions and structural invariants can be used to change the internal properties of the system when the system is fully actuated, and how it can be done in an approximate way when the system is actuated using piecewise continuous actuators stemming from the use of patches. Asymptotic stability is achieved using damping injection. In the boundary-controlled case we show how the closed loop energy function can be partially shaped, modifying the minimum and a part of the shape of this function and how damping injection can be used to guarantee asymptotic convergence. We end with some some extensions of the proposed results to irreversible thermodynamic systems.

- Masaaki Nagahara (The University of Kitakyushu, Japan)

Compressed sensing and maximum hands-off control

Abstract. Compressed sensing has been actively researched in the field of signal processing and machine learning. More recently, the method has been applied to control problems. In this talk, we will briefly review compressed sensing for vectors, and then introduce the maximum hands-off control for continuous-time systems, which aims at finding the sparsest control under control constraints.

Full paper *Masaaki Nagahara* "Compressed sensing and maximum hands-off control" available in the IFAC PapersOnline volume of MTNS 2022 https://doi.org/10.1016/j.ifacol.2022.11.097.

- Na Li (Harvard University, Cambridge, MA, USA)

Scalable distributed control and learning of networked dynamical systems

Abstract. Recent radical evolution in distributed sensing, computation, communication, and actuation has fostered the emergence of cyber-physical network systems. Regardless of the specific application, one central goal is to shape the network collective behavior through the design of admissible local decision-making algorithms. This is nontrivial due to various challenges such as the local connectivity, system complexity and uncertainty, limited information structure, and the complex intertwined physics and human interactions.

In this talk, I will present our recent progress in formally advancing the systematic design of distributed coordination in network systems via harnessing special properties of the underlying problems and systems. In particular, we will present three examples and discuss three type of properties, i) how to use network structure to ensure the performance of the local controllers; ii) how to use the information and communication structure to develop distributed learning rules; iii) how to use domain-specific properties to further improve the efficiency of the distributed control and learning algorithms.

- Jacquelien M. A. Scherpen (University of Groningen, The Netherlands)

Extended (differential) balancing for model reduction of linear and nonlinear dynamical systems

Abstract. In this talk, we will develop extended balancing and its structure preservation possibilities for linear systems, as well as extended balancing theory for nonlinear systems in the contraction framework. For the latter, we introduce the concept of the extended differential observability Gramian and inverse of the extended differential controllability Gramian for nonlinear dynamical systems and show their correspondence with generalized differential Gramians. We also provide how extended (differential) balancing can be utilized for model reduction to get a smaller apriori error bound in comparison with generalized (differential balancing). We will focus on preserving the structure of a port-Hamiltonian system with help of extended balancing in both the linear and nonlinear systems setting.

- Sanne ter Horst (North-West University, Potchefstroom, South Africa)

Convex invertible cones and Nevanlinna-Pick interpolation

Abstract. Nevanlinna-Pick interpolation developed from a topic in classical complex analysis to a useful tool for solving various problems in control theory and electrical engineering. Over the years many extensions of the original problem were considered, including extensions to different function spaces, nonstationary problems, several variable settings and interpolation with matrix and operator points. In this talk we discuss a variation on Nevanlinna-Pick interpolation for positive real odd functions evaluated in real matrix points. This problem was studied by Cohen and Lewkowicz using convex invertible cones and the Lyapunov order, making interesting connections with stability theory. The solution requires an analysis of linear matrix maps using representations that go back to work of R.D. Hill from the 1970s and focusses, in particular, on the question when positive linear matrix maps are completely positive. If time permits, some possible extensions to multidimensional systems will briefly be discussed.

Full paper *S. ter Horst, A. van der Merwe* "Convex invertible cones and Nevanlinna-Pick interpolation: The suboptimal case" available in the IFAC PapersOnline volume of MTNS 2022 https://doi.org/10.1016/j.ifacol.2022.11.050.

- Claudia Schillings (Free University Berlin, Germany)

A General Framework for Machine Learning-based Optimization Under Uncertainty

Abstract. Approaches to decision making and learning mainly rely on optimization techniques to achieve “best” values for parameters and decision variables. In most practical settings, however, the optimization takes place in the presence of uncertainty about model correctness, data relevance, and numerous other factors that influence the resulting solutions. For complex processes modeled by nonlinear ordinary and partial differential equations, the incorporation of these uncertainties typically results in high or even infinite dimensional problems in terms of the uncertain parameters as well as the optimization variables, which in many cases are not solvable with current state of the art methods. One promising potential remedy to this issue lies in the approximation of the forward problems using novel techniques arising in uncertainty quantification and machine learning.

We propose in this talk a general framework for machine learning based optimization under uncertainty and inverse problems. Our approach replaces the complex forward model by a surrogate, e.g. a neural network, which is learned simultaneously in a one-shot sense when estimating the unknown parameters from data or solving the optimal control problem. By establishing a link to the Bayesian approach, an algorithmic framework is developed which ensures the feasibility of the parameter estimate / control w.r. to the forward model.

This is joint work with Philipp Guth (U Mannheim) and Simon Weissmann (U Heidelberg).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Stability of solutions for controlled nonlinear systems under perturbation of state constraints

Pierre-Cyril Aubin-Frankowski *

* *INRIA Paris, France (e-mail: pierre-cyril.aubin@inria.fr).*

Abstract: This paper tackles the problem of nonlinear systems, with sublinear growth but unbounded control, under perturbation of some time-varying state constraints. It is shown that, given a trajectory to be approximated, one can find a neighboring one that lies in the interior of the constraints, and which can be made arbitrarily close to the reference trajectory both in L^∞-distance and L^2-control cost. This result is an important tool to prove the convergence of approximation schemes of state constraints based on interior solutions and is applicable to control-affine systems.

Keywords: Nonlinear control systems, Control of constrained systems, Time-varying systems, Interior trajectories

1. INTRODUCTION

We consider a nonlinear system with unbounded control and state constraints

$$\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \qquad \text{for a.e. } t \in [0,T], \quad (1)$$
$$\mathbf{x}(t) \in \mathcal{A}_{0,t} := \{\mathbf{x} \,|\, \mathbf{h}(t,\mathbf{x}) \leq 0\}, \qquad \text{for all } t \in [0,T], \quad (2)$$

where $\mathbf{f} : [0,T] \times \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ and $\mathbf{h} : [0,T] \times \mathbb{R}^N \to \mathbb{R}^P$. Given a reference trajectory $\bar{\mathbf{x}}(\cdot)$, such that $\mathbf{h}(0, \bar{\mathbf{x}}(0)) < 0$, with control $\bar{\mathbf{u}}(\cdot)$ satisfying (1)-(2), our goal is to design a trajectory $\mathbf{x}^\epsilon(\cdot)$ with the same initial condition and some control $\mathbf{u}^\epsilon(\cdot)$, chosen such that $\mathbf{x}^\epsilon(\cdot)$ can be made arbitrarily L^∞-close to $\bar{\mathbf{x}}(\cdot)$, with $\mathbf{u}^\epsilon(\cdot)$ having almost the same L^2-norm as $\bar{\mathbf{u}}(\cdot)$, while also satisfying (1) and the following tightened constraints:

$$\mathbf{x}^\epsilon(t) \in \mathcal{A}_{\epsilon,t} := \{\mathbf{x} \,|\, \boldsymbol{\epsilon} + \mathbf{h}(t,\mathbf{x}) \leq 0\} \text{ for all } t \in [0,T]. \quad (3)$$

This construction is crucial to prove the convergence of approximation schemes of the constraints from within, in the sense that if $(\bar{\mathbf{x}}(\cdot), \bar{\mathbf{u}}(\cdot))$ is the solution of some optimal control problem with quadratic cost in control, then $(\mathbf{x}^\epsilon(\cdot), \mathbf{u}^\epsilon(\cdot))$ would be almost optimal while strictly interior. Such schemes were used by the author in Aubin-Frankowski (2021) for linear $\mathbf{f}$ and $\mathbf{h}$, leveraging convexity of the set of trajectories for which (1)-(2) hold. Here we provide instead assumptions on $\mathbf{f}$ and $\mathbf{h}$ designed originally by Bettiol et al. (2012) for bounded differential inclusions with time-invariant constraints. We further improve on their construction to have both an estimate on the L^2-norm and to cover unbounded systems (1) and time-varying constraints (2). This analysis can be related also to Bettiol and Vinter (2011) where time-dependent bounded systems are considered. The prototypical cases we are interested in are constrained nonlinear control-affine systems as studied e.g. in a more restrictive setting in (Cannarsa and Castelpietra, 2008, Section 4).

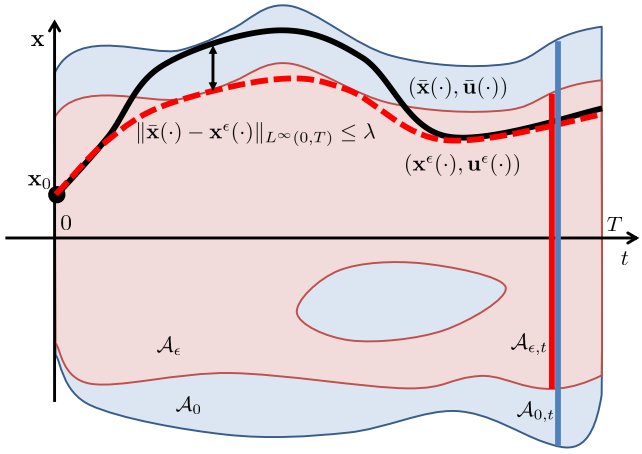


Figure 1. Illustration of the trajectories and constraints considered in Theorem 1.

2. MAIN RESULT

Notations. The integer interval is written $[\![i,j]\!] = \{i, i+1, \dots, j\}$. We denote by $\mathbb{R}_+$ the set of nonnegative reals, and use the shorthand $L^p(0,T)$ for $L^p([0,T],\mathbb{R}^d)$ with $p \in \{1,2,\infty\}$, and L^p_+ when the output set is $\mathbb{R}^d_+$. The set $\mathbb{B}_d$ is the closed Euclidean unit ball of $\mathbb{R}^d$ of center $\mathbf{0}$. Given a set $\mathcal{A} \subset \mathbb{R}^d$, $\mathrm{Int}(\mathcal{A})$ designates its interior, $\partial\mathcal{A}$ its boundary, and $d_\mathcal{A}(\cdot)$ is the Euclidean distance to $\mathcal{A}$.

We call $\mathbf{f}$-trajectories the solutions of (1) for measurable controls $\mathbf{u}(\cdot)$. For any $\boldsymbol{\epsilon} \in \mathbb{R}^P_+$, define

$$\mathcal{A}_\epsilon := \{(t,\mathbf{x}) \,|\, t \in [0,T],\, \mathbf{x} \in \mathcal{A}_{\epsilon,t}\}.$$

A trajectory is said to be $\mathcal{A}_\epsilon$-feasible if (3) holds, for instance $\bar{\mathbf{x}}(\cdot)$ is $\mathcal{A}_0$-feasible by assumption. We define the maximal constraint violation $\rho_{\epsilon,[t_0,t_1]}(\mathbf{x}(\cdot))$ of a trajectory on an interval $[t_0,t_1] \subset [0,T]$ as follows

$$\rho_{\epsilon,[t_0,t_1]}(\mathbf{x}(\cdot)) := \sup_{t\in[t_0,t_1]} d_{\mathcal{A}_{\epsilon,t}}(\mathbf{x}(t)).$$

We assume from now on that the control $\bar{\mathbf{u}}(\cdot)$ belongs to $L^\infty(0,T)$. This restriction is due to (H-6) below since we use a time-delay in the construction of the control $\mathbf{u}^\epsilon(\cdot)$ that is ill-suited to track the distance between controls. If $\mathbf{f}(t,\cdot,\mathbf{u})$ is $k_f(t)$-Lipschitz for any $\mathbf{u}\in\mathbb{R}^M$, then we may just assume that $\bar{\mathbf{u}}(\cdot)\in L^2(0,T)$.

(H-1) (Regular perturbation of $\mathcal{A}$)
$$\forall\,\lambda>0,\ \exists\,\boldsymbol{\epsilon}>0,$$
$$\forall\,(t,\mathbf{x})\in\mathcal{A}_0\cap([0,T]\times\|\bar{\mathbf{x}}(\cdot)\|_{L^\infty(0,T)}\mathbb{B}_N),$$
$$d_{\mathcal{A}_{\epsilon,t}}(\mathbf{x})\le\lambda.$$

(H-2) (Uniform continuity from the right of $d_{\partial\mathcal{A}_{\epsilon,t}}$ w.r.t. ϵ and t) There exist $\boldsymbol{\epsilon}_0>0$, $\Delta_0>0$, and a continuous function $\omega_{\mathcal{A}}(\cdot)\in\mathcal{C}^0(\mathbb{R}_+,\mathbb{R}_+)$ such that $\omega_{\mathcal{A}}(0)=0$ and, for all $\boldsymbol{\epsilon}\le\boldsymbol{\epsilon}_0$, and all $(t,\mathbf{x})\in\mathcal{A}_0\cap([0,T]\times 2\|\bar{\mathbf{x}}(\cdot)\|_{L^\infty(0,T)}\mathbb{B}_N)$,
$$\forall\,\delta\in[0,\min(\Delta_0,T-t)],\ \|d_{\partial\mathcal{A}_{\epsilon,t+\delta}}(\mathbf{x})-d_{\partial\mathcal{A}_{\epsilon,t}}(\mathbf{x})\|\le\omega_{\mathcal{A}}(\delta).$$

(H-3) (Sublinear growth of $\mathbf{f}$ w.r.t. $\mathbf{x}$ and $\mathbf{u}$)
$$\exists\,\theta(\cdot)\in L^2_+(0,T),\ \forall\,t\in[0,T],\ \forall\,\mathbf{x}\in\mathbb{R}^N,\ \forall\,\mathbf{u}\in\mathbb{R}^M,$$
$$\|\mathbf{f}(t,\mathbf{x},\mathbf{u})\|\le\theta(t)(1+\|\mathbf{x}\|+\|\mathbf{u}\|).$$

(H-4) (Inward-pointing condition) There exist $\boldsymbol{\epsilon}_0>0$, $M_u>0$, $M_v>0$, $\xi>0$, and $\eta>0$ such that for all $\boldsymbol{\epsilon}\le\boldsymbol{\epsilon}_0$ and all $(t,\mathbf{x})\in(\partial\mathcal{A}_\epsilon+(0,\eta\mathbb{B}_N))\cap\mathcal{A}_\epsilon\cap([0,T]\times(1+2\|\bar{\mathbf{x}}(\cdot)\|_{L^\infty(0,T)})\mathbb{B}_N)$, we can find $\mathbf{u}\in M_u\mathbb{B}_M$ such that $\mathbf{v}:=\mathbf{f}(t,\mathbf{x},\mathbf{u})$ belongs to $M_v\mathbb{B}_N$ and
$$\mathbf{y}+\delta(\mathbf{v}+\xi\mathbb{B}_N)\subset\mathcal{A}_{\epsilon,t+\delta}\tag{4}$$
for all $\delta\in[0,\xi]$ and all $\mathbf{y}\in(\mathbf{x}+\xi\mathbb{B}_N)\cap\mathcal{A}_{\epsilon,t}$.

(H-5) (Left local absolute continuity of $\mathbf{f}$ w.r.t. t)
$$\exists\,\gamma(\cdot)\in L^1_+(0,T),\ \exists\,\beta_u(\cdot)\in L^2_+(0,T),$$
$$\forall\,0\le s<t\le T,\ \forall\,\mathbf{x}\in(1+2\|\bar{\mathbf{x}}(\cdot)\|_{L^\infty(0,T)})\mathbb{B}_N,$$
$$\forall\,\mathbf{u}_s\in(M_u+\|\bar{\mathbf{u}}(s)\|)\mathbb{B}_M,\ \exists\,\mathbf{u}_t\in\mathbf{u}_s+\beta_u(s)\mathbb{B}_M,$$
$$\|\mathbf{f}(t,\mathbf{x},\mathbf{u}_t)-\mathbf{f}(s,\mathbf{x},\mathbf{u}_s)\|\le\int_s^t\gamma(\sigma)\,d\sigma.$$

Let $R:=e^{\|\theta(\cdot)\|_{L^1(0,T)}}\big[1+\|\bar{\mathbf{x}}(\cdot)\|_{L^\infty(0,T)}$
$+(1+M_u)\|\theta(\cdot)\|_{L^1(0,T)}$
$+\|\theta(\cdot)\|_{L^2(0,T)}(\|\bar{\mathbf{u}}(\cdot)\|_{L^2(0,T)}+\|\beta_u(\cdot)\|_{L^2(0,T)})\big]$. (5)

(H-6) (Local Lipschitz continuity of $\mathbf{f}$ w.r.t. $\mathbf{x}$)
$$\exists\,k_f(\cdot)\in L^2_+(0,T),\ \forall\,t\in[0,T],\ \forall\,\mathbf{x},\mathbf{y}\in R\mathbb{B}_N,$$
$$\forall\,\mathbf{u}\in(M_u+\|\bar{\mathbf{u}}(\cdot)\|_{L^\infty(0,T)})\mathbb{B}_M,$$
$$\|\mathbf{f}(t,\mathbf{x},\mathbf{u})-\mathbf{f}(t,\mathbf{y},\mathbf{u})\|\le k_f(t)\|\mathbf{x}-\mathbf{y}\|.$$

(H-7) (Hölderian selection of the controls in (H-5))
$$\exists\,\gamma(\cdot)\in L^1_+(0,T),\ \exists\,\alpha\in]0,1],\ \exists\,k_u(\cdot)\in L^2_+(0,T),$$
$$\forall\,0\le s<t\le T,\ \forall\,\mathbf{x}\in(1+2\|\bar{\mathbf{x}}(\cdot)\|_{L^\infty(0,T)})\mathbb{B}_N,$$
$$\forall\,\mathbf{u}_s\in(M_u+\|\bar{\mathbf{u}}(s)\|)\mathbb{B}_M,\ \exists\,\mathbf{u}_t\in\mathbf{u}_s+(t-s)^\alpha k_u(s)\mathbb{B}_M,$$
$$\|\mathbf{f}(t,\mathbf{x},\mathbf{u}_t)-\mathbf{f}(s,\mathbf{x},\mathbf{u}_s)\|\le\int_s^t\gamma(\sigma)\,d\sigma.$$

Theorem 1. Under assumptions (H-1)-(H-6), for any $\lambda>0$, there exists $\boldsymbol{\epsilon}>0$ and a $\mathbf{f}$-trajectory $\mathbf{x}^\epsilon(\cdot)$ on $[0,T]$ such that $\mathbf{x}^\epsilon(0)=\bar{\mathbf{x}}(0)$, $\mathbf{x}^\epsilon(t)\in\operatorname{Int}\mathcal{A}_{\epsilon,t}$ for all $t\in[0,T]$, and
$$\|\bar{\mathbf{x}}(\cdot)-\mathbf{x}^\epsilon(\cdot)\|_{L^\infty(0,T)}\le\lambda.$$
Moreover if (H-7) is satisfied, then, for any mapping $\mathbf{R}(\cdot)\in\mathcal{C}^0([0,T],\mathbb{R}^{M,M})$ with positive semidefinite matrix values, one can choose $\boldsymbol{\epsilon}>0$ and $\mathbf{x}^\epsilon(\cdot)$ such that the controls $\mathbf{u}^\epsilon(\cdot)$ satisfy
$$\left|\|\mathbf{R}(\cdot)^{1/2}\bar{\mathbf{u}}(\cdot)\|^2_{L^2(0,T)}-\|\mathbf{R}(\cdot)^{1/2}\mathbf{u}^\epsilon(\cdot)\|^2_{L^2(0,T)}\right|\le\lambda.$$

Discussion of the assumptions: The properties (H-1) and (H-2) imposed on the constraint set can for instance be derived from the $\mathcal{C}^{1,1}$-regularity of $\mathbf{h}$, coupled with the assumptions that the Jacobian $\frac{\partial\mathbf{h}(t,\mathbf{x})}{\partial\mathbf{x}}$ of $\mathbf{h}$ at all $(t,\mathbf{x})\in\partial\mathcal{A}_0$ is surjective. The classical growth assumption (H-3) prevents in turn finite-time explosion of the trajectories. The Lipschitzianity (H-6) guarantees their uniqueness and was designed to encompass control-affine systems of the form $\mathbf{x}'(t)=\mathbf{a}(t,\mathbf{x})+\mathbf{b}(t,\mathbf{x})\mathbf{u}$ with $\tilde{k}_f(t)$-Lipschitz functions $\mathbf{a}(t,\cdot)$ and $\mathbf{b}(t,\cdot)$, for some $\tilde{k}_f(\cdot)\in L^2(0,T)$. The other assumptions are more technical and inspired by Bettiol et al. (2012). Inward-pointing conditions such as (H-4), which can be deduced from a normal cone formulation (Bettiol et al., 2012, Lemma 5.3), have been shown to yield the L^∞-bounds we seek. The time regularity (H-5) was introduced to tackle discontinuities in the dynamics, and showcased on a civil engineering example (Bettiol et al., 2012, Section 4). We adapt it to control systems and refine it in (H-7).

Note that state constraints of order 2 (or more), e.g. $\ddot{x}=u$ with x constrained, do not enter into the proposed framework as the inward-pointing assumption does not hold in these cases, being limited to "order 1 constraints". The proof can in principle be adapted to systems $\tilde{\mathbf{f}}$ with control constraints following a Lipschitz (or Hölderian) closed-valued map $t\rightsquigarrow U(t)$ by considering the projection over $U(t)$ and $\mathbf{f}(t,\mathbf{x},\mathbf{u})=\tilde{\mathbf{f}}(t,\mathbf{x},\operatorname{proj}_{U(t)}(\mathbf{u}))$ assuming that this $\mathbf{f}$ satisfies the above assumptions.

Idea of the proof: The overall strategy to construct a neighboring $\mathcal{A}_\epsilon$-feasible trajectory can be related to that of Bettiol et al. (2012). Modifying it to unbounded controls and time-varying constraints is however not straightforward. We start by considering small subintervals $[0,T]=\bigcup_{i\in[\![0,N_0-1]\!]}[t_i,t_{i+1}]$ and proceed iteratively. If the ith-trajectory stays in $\mathcal{A}_\epsilon$ over $[t_i,t_{i+1}]$, we move to the next time interval. Otherwise for the $(i+1)$th-trajectory over $[t_i,t_{i+1}]$, (H-4) provides us with an inward-pointing control to stay in $\mathcal{A}_\epsilon$ for a short time. Then we apply a delayed control of the ith-trajectory for the rest of $[t_i,t_{i+1}]$, and the original control $\bar{\mathbf{u}}(\cdot)$ over $[t_{i+1},T]$. By monitoring several quantities, we can show that the resulting control after N_0 iterations is L^2-close from $\bar{\mathbf{u}}(\cdot)$ and that the obtained trajectory is in $\mathcal{A}_\epsilon$.

Example: Consider an electric motor
$$x'(t)=a(t,x)+b(t,u),$$
with a bounded $a\in C^{1,1}([0,2]\times\mathbb{R},\mathbb{R})$ and constraints $h(x)=1-|x|$, for controls $u\in\mathbb{R}$. The motor suffers an incident at $T=1$. If it is a power surge

$$b(t,u) = \tilde{b}(t)u = \begin{cases} u & \text{if } t \in [0,1] \\ u/\sqrt[4]{t-1} & \text{if } t \in]1,2] \end{cases},$$

then (H-3) holds for $\theta \equiv \tilde{b} + \|f\|_\infty$. It remains to check (H-7). For $s < t \leq 1$, take $u_t = u_s$. For $s < 1 < t$, set $u_t = \sqrt[4]{t-1}u_s$, so $|u_t - u_s| \leq |u_s| \leq \frac{\sqrt[4]{t-s}}{\sqrt[4]{|1-s|}}|u_s|$. For $s = 1 < t$, take also $u_t = \sqrt[4]{t-1}u_s$, thus $|u_t - u_s| \leq |u_s|$. For $1 < s \leq t$, set $u_t = \frac{\sqrt[4]{t-1}}{\sqrt[4]{s-1}}u_s$, so, by subadditivity of $\sqrt[4]{\cdot}$,

$$|u_t - u_s| \leq \frac{\sqrt[4]{t-1} - \sqrt[4]{s-1}}{\sqrt[4]{|1-s|}}|u_s| \leq \frac{\sqrt[4]{t-s}}{\sqrt[4]{|1-s|}}|u_s|.$$

Hence (H-7) is indeed satisfied for $\gamma \equiv 0$, $k_u(s) = \frac{1}{\sqrt[4]{|1-s|}}(M_u + |\bar{\mathbf{u}}(s)|)$ and $\alpha = \frac{1}{4}$. If the incident consists in a power decline

$$b(t,u) = \begin{cases} \arctan(u) & \text{if } t \in [0,1] \\ (1 - \frac{\sqrt{t-1}}{2})\arctan(u) & \text{if } t \in]1,2] \end{cases},$$

then the system is bounded and (H-7) holds with $u_t = u_s$, $\gamma(\sigma) = \frac{1}{4\sqrt{\sigma-1}}$ for $\sigma \in]1,2]$ and $\gamma(\sigma) = 0$ otherwise. In both cases (H-4) is satisfied, so perturbing the constraints still allows for a trajectory and control close to the reference ones as per Theorem 1.

REFERENCES

Aubin-Frankowski, P.C. (2021). Linearly constrained linear quadratic regulator from the viewpoint of kernel methods. *SIAM Journal on Control and Optimization*, 59(4), 2693–2716. doi:10.1137/20m1348765.

Bettiol, P., Frankowska, H., and Vinter, R.B. (2012). L^∞ estimates on trajectories confined to a closed subset. *Journal of Differential Equations*, 252, 1912–1933. doi:10.1016/j.jde.2011.09.007.

Bettiol, P. and Vinter, R. (2011). Trajectories satisfying a state constraint: Improved estimates and new nondegeneracy conditions. *IEEE Transactions on Automatic Control*, 56(5), 1090–1096. doi:10.1109/tac.2010.2088670.

Cannarsa, P. and Castelpietra, M. (2008). Lipschitz continuity and local semiconcavity for exit time problems with state constraints. *Journal of Differential Equations*, 245(3), 616–636. doi:10.1016/j.jde.2007.10.020.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Separating invariants for matrix tuples up to similarity

J. Volčič *

* *Department of Mathematics, Drexel University, PA, USA*

Abstract: The talk considers evaluations of linear matrix pencils $L = T_0 + x_1T_1 + \cdots + x_mT_m$ on matrix tuples as $L(X_1, \ldots, X_m) = I \otimes T_0 + X_1 \otimes T_1 + \cdots + X_m \otimes T_m$. It is shown that ranks of linear matrix pencils constitute a collection of separating invariants for simultaneous similarity of matrix tuples. That is, m-tuples X and Y of $n \times n$ matrices are simultaneously similar if and only if $\mathrm{rk}L(X) = \mathrm{rk}L(Y)$ for all linear matrix pencils L of size mn. Variants of this property for some other group actions are also discussed.

Keywords: Simultaneous similarity, linear matrix pencil, rank-preserving map.

Two tuples of $n \times n$ matrices $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_m)$ over a field are *(simultaneously) similar* if there exists $P \in \mathrm{GL}_n$ such that $Y_i = PX_iP^{-1}$ for $i = 1, \ldots, m$. The classification of matrix tuples up to similarity has been deemed a "hopeless problem". Nevertheless, the study of simultaneous similarity and related group actions on matrix tuples is crucial in multiple areas of mathematics, ranging from operator theory, invariant and representation theory and algebraic geometry to algebraic statistics and computational complexity. A prominent facet of simultaneous similarity is finding a (natural) collection of separating invariants, which is the topic of this talk. A related problem is that of the orbit closure inclusion: determine whether a matrix tuple X belongs to the closure of the similarity orbit of a matrix tuple Y. This problem is fundamental to geometric invariant theory and geometric complexity theory. Note that X and Y are similar if and only if X is in the orbit closure of Y and Y is in the orbit closure of X.

Curto and Herrero (1985) conjectured that X lies in the closure of the similarity orbit of Y if and only if $\mathrm{rk}f(X) \leq \mathrm{rk}f(Y)$ for every noncommutative polynomial f in m variables. Hadwin and Larson (2003) gave a counterexample to the (even weaker) two-sided Curto–Herrero conjecture: they presented matrix tuples x and Y that are not similar but $\mathrm{rk}f(X) = \mathrm{rk}f(Y)$ for every noncommutative polynomial f. Furthermore, they proposed an ameliorated conjecture: X lies in the closure of the similarity orbit of Y if and only if $\mathrm{rk}F(X) \leq \mathrm{rk}F(Y)$ for every *matrix noncommutative polynomial* F (i.e., a matrix of noncommutative polynomials).

This talk presents the affirmative answer to two-sided version of the Hadwin–Larson conjecture.

Theorem 1. [Derksen et al. (2021)] The following are equivalent for $X, Y \in \mathrm{Mat}_n^m$:

(1) X and Y are similar;
(2) for every $T = (T_0, \ldots, T_m) \in \mathrm{Mat}_{mn}^{m+1}$,

$$\begin{aligned} &\mathrm{rk}\,(I \otimes T_0 + X_1 \otimes T_1 + \cdots X_m \otimes T_m) \\ &= \mathrm{rk}\,(I \otimes T_0 + Y_1 \otimes T_1 + \cdots Y_m \otimes T_m)\,. \end{aligned}$$

In other words, ranks of linear matrix pencils evaluated at matrix tuples constitute a collection of separating invariants for simultaneous similarity. Similar results hold for the actions of unitary, orthogonal and symplectic groups, and for the left-right action of the general linear group.

Moreover, the talk provides a counterexample to the general version of the Hadwin–Larson conjecture: there are two pairs X and Y of 4×4 matrices such that $\mathrm{rk}F(X) \leq \mathrm{rk}F(Y)$ for all matrix noncommutative polynomials F but X does not lie the closure of the similarity orbit of Y.

This talk is based on joint work with Harm Derksen, Igor Klep and Visu Makam.

REFERENCES

Curto, R.E. and Herrero, D.A. (1985). On closures of joint similarity orbits. *Integral Equations Operator Theory*, 8, 489–556.

Derksen, H., Klep, I., Makam, V., and Volčič, J. (2021). Ranks of linear matrix pencils separate simultaneous similarity orbits. *preprint arXiv:2109.09418*.

Hadwin, D. and Larson, D.R. (2003). Completely rank-nonincreasing linear maps. *J. Funct. Anal.*, 199, 210–227.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Globally positive trace polynomials

J. Volčič *

* *Department of Mathematics, Drexel University, PA, USA*

Abstract: A trace polynomial is a polynomial in noncommuting variables and traces of their products. It is positive if its evaluations on all symmetric matrices, or more generally, self-adjoint operators from tracial von Neumann algebras, attain only positive semidefinite values. A Positivstellensatz for positive univariate trace polynomials is presented, and a characterization of trace-positive multivariate noncommutative polynomials is discussed.

Keywords: Trace polynomial, Positivstellensatz, moment problem.

Trace polynomials are real polynomials in noncommuting variables $x_1, \dots, x_d$ and their formal traces $\mathrm{tr}(x_{i_1} \dots x_{i_\ell})$. Such expressions can be naturally evaluated on tuples of matrices, where the trace symbols are evaluated as normalized traces; or more generally, on tuples of operators from a tracial von Neumann algebra. Trace polynomials as matricial/operator functions originated in invariant theory, and more recently emerged in free probability and quantum information theory.

This talk discusses positive trace polynomials, i.e., those that can attain only positive semidefinite values when evaluated on tuples of symmetric matrices, or more generally self-adjoint operators. This topic is well understood when evaluations are restricted to either matrix tuples of a fixed dimension (Klep et al. (2018)) or bounded domains in tracial von Neumann algebras (Klep et al. (2022)). On the other hand, results are very scarce in the global case, when restrictions on dimensions and boundedness are dropped.

The majority of the talk is dedicated to the univariate case ($d = 1$ and $x_1 = x$). Univariate trace polynomials form a commutative polynomial ring (in countably many variables), and several sum-of-squares positivity certificates (Positivstellensätze) in commutative rings are provided by real algebraic geometry. However, this theory does not appear to directly apply to our setup. First, matrix evaluations of trace polynomials are just a special class of homomorphisms on trace polynomials. Second, the dimension-free context addresses positivity on matrices of all sizes, hence on a countable disjoint union of real affine spaces; there is no bound (with respect to the degree of a univariate trace polynomial) on the size of matrices for which positivity needs to be verified.

Therefore a different approach is required. To demonstrate it, consider the inequality

$$\begin{aligned} &\mathrm{tr}(X^4)\mathrm{tr}(X^2) + 2\mathrm{tr}(X^3)\mathrm{tr}(X^2)\mathrm{tr}(X) \\ \geq\ &\mathrm{tr}(X^4)\mathrm{tr}(X)^2 + \mathrm{tr}(X^3)^2 + \mathrm{tr}(X^2)^3 \end{aligned} \tag{1}$$

for all symmetric matrices X (and normalized trace tr). One way to certify (1) is by noticing that the trace polynomial $f = \mathrm{tr}(x^4)(\mathrm{tr}(x^2) - \mathrm{tr}(x)^2) + 2\mathrm{tr}(x^3)\mathrm{tr}(x^2)\mathrm{tr}(x) - \mathrm{tr}(x^3)^2 - \mathrm{tr}(x^2)^3$ satisfies

$$\begin{aligned} &\mathrm{tr}\Big(\Big(x - \mathrm{tr}(x)\Big)^2\Big) \cdot f \\ =\ &\mathrm{tr}\Big(\Big((\mathrm{tr}(x)^2 - \mathrm{tr}(x^2))x^2 + (\mathrm{tr}(x^3) - \mathrm{tr}(x^2)\mathrm{tr}(x))x \\ &\quad + \mathrm{tr}(x^2)^2 - \mathrm{tr}(x^3)\mathrm{tr}(x)\Big)^2\Big), \end{aligned}$$

where we view tr as an idempotent linear endomorphism of trace polynomials in a natural way. More generally, the following tracial analog of Artin's solution to Hilbert's 17th problem holds.

Theorem 1. [Klep et al. (2021)] A univariate trace polynomial is positive on all symmetric matrices if and only if it is a quotient of sums of products of traces of squares of trace polynomials.

It turns out that in the multivariate case ($d > 1$), traces of squares of polynomials are are not sufficient for describing positivity. For example, the tracial analog of Motzkin's example,

$$\mathrm{tr}(x_1x_2^4x_1 + x_2x_1^4x_2 - 3x_1x_2^2x_1 + 1)$$

is positive for all pairs of symmetric matrices, but cannot be described by traces of squares of noncommutatve polynomials. Nevertheless, a characterization of trace-positive noncommutative polynomials in terms of noncommutative rational functions is given, based on a solution of the unbounded tracial moment problem.

This talk is based on joint work with Igor Klep and James Pascoe.

REFERENCES

Klep, I., Magron, V., and Volčič, J. (2022). Optimization over trace polynomials. *Ann. Henri Poincaré*, 23, 67–100.

Klep, I., Pascoe, J.E., and Volčič, J. (2021). Positive univariate trace polynomials. *J. Algebra*, 579, 303–317.

Klep, I., Š. Špenko, and Volčič, J. (2018). Positive trace polynomials and the universal procesi-schacher conjecture. *Proc. London Math. Soc.*, 117, 1101–1134.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Bézout Identity in Pseudoratoinal Transfer Functions

Extended abstract for MTNS 2022

Yutaka Yamamoto [*,1]

[*] *Professor Emeritus, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: yy@i.kyoto-u.ac.jp).*

Catherine Bonnet [**]

[**] *Inria, Université Paris-Saclay, L2S–CentraleSupélec, 3 rue Joliot Curie 91192 Gif-sur-Yvette cedex France.* Catherine.Bonnet@inria.fr.

Abstract: Coprime factorizations of transfer functions play various important roles, e.g., minimality of realizations, stabilizability of systems, etc. This paper studies the Bézout condition over the ring $\mathscr{E}'(\mathbb{R}_-)$ of distributions of compact support and the ring $\mathfrak{M}(\mathbb{R}_-)$ of measures with compact support. These spaces are known to play crucial roles in minimality of state space representations and controllability of behaviors. We give a detailed review of the results obtained thus far, as well as discussions on a new attempt of deriving general results from that for measures. It is clarified that there is a technical gap in generalizing the result for $\mathfrak{M}(\mathbb{R}_-)$ to that for $\mathscr{E}'(\mathbb{R}_-)$. A detailed study of a concrete example is given.

Keywords: Bézout identity, pseudorationality, distributions, Gel'fand representation, delay-differential systems
AMS subject classification: 46F10, 46J15

1. INTRODUCTION

This short note studies the issue of coprimeness for a certain class of infinite-dimensional systems.

In particular, we study the Bézout identity (or Bézout condition)

$$px+qy=1 \tag{1}$$

in an algebra appropriate for a class of distributed parameter systems.

The first author has introduced the class of *pseudorational* impulse responses or transfer functions, and developed realization theory, various spectral analysis, and coprimeness conditions Yamamoto (1988, 1989). The present article is an extended abstract version of the paper for MTNS 2020. For more background explanations, we refer the reader to Yamamoto and Bonnet (2021).

2. PSEUDORATIONALITY

Let $\mathscr{D}'$ denotes the space of distributions on $\mathbb{R}$. Let $\mathscr{E}'(\mathbb{R})$ be its subspace consisting of those having compact support. $\mathscr{E}'(\mathbb{R}_-)$ is also its subspace with support contained in the negative half line $(-\infty,0]$. $\mathscr{D}'_+$ denotes the subspace of $\mathscr{D}'$ consisting of elements having support bounded on the left. Distributions such as Dirac's delta δ_a placed at $a\in\mathbb{R}$, its derivative δ'_a are examples of elements in $\mathscr{E}'(\mathbb{R})$. If $a\le 0$, then they belong to $\mathscr{E}'(\mathbb{R}_-)$.

We consider fraction representations over $\mathscr{E}'(\mathbb{R}_-)$.

Definition 2.1. An impulse response function G ($\operatorname{supp} G\subset[0,\infty)$) is said to be *pseudorational* (Yamamoto (1988)) if there exist $q,p\in\mathscr{E}'(\mathbb{R}_-)$ such that

(1) $G=q^{-1}*p$ where the inverse is taken with respect to convolution and belongs to $\mathscr{D}'_+$;
(2) $\operatorname{ord} q^{-1}=-\operatorname{ord} q$, where $\operatorname{ord} q$ denotes the order of a distribution q (Schwartz (1966)). [2]

If this condition is satisfied, we call (p,q) a *pseudorational pair*. The Laplace transform $\hat{q}^{-1}\hat{p}$ is called a *pseudorational transfer function*.

The delay-differential equation:

$$\dot{x}(t)=x(t-1)+u(t)$$
$$y(t)=x(t),$$

admits the representation

$$y=(\delta'_{-1}-\delta)^{-1}*\delta_{-1}*u,$$

and hence it is pseudorational.

[⋆] The present article is an extended abstract version of the paper Yamamoto and Bonnet (2021) for MTNS 2020.

[1] This author was supported in part by the Japan Society for the Promotion of Science under Grants-in-Aid for Scientific Research No. 19H02161. The author also wishes to thank DIGITEO and Laboratoire des Signaux et Systemes (L2S, UMR CNRS), CNRS-CentraleSupelec-University Paris-Sud and Inria Saclay for their financial support while part of this research was conducted.

[2] Roughly speaking, the order of a distribution α is the least integer r such that $\alpha=(d/dt)^r\beta$ for some measure β.

The main problem that concerns us here is the following:
Problem *Given a pseudorational pair* $(p,q) \in \mathscr{E}'(\mathbb{R}_-) \times \mathscr{E}'(\mathbb{R}_-)$*, characterize a condition under which p and q satisfy the Bézout identity:*

$$p*x+q*y=\delta \tag{2}$$

for some $x,y \in \mathscr{E}'(\mathbb{R}_-)$.

If we consider $\mathscr{E}'(\mathbb{R})$ instead of $\mathscr{E}'(\mathbb{R}_-)$, it gives a necessary and sufficient condition for the controllability of the behavior defined over $\mathscr{D}'$ (Yamamoto (2016)). Actually, the Bézout condition over $\mathscr{E}'(\mathbb{R})$ is in close relationship with that in $\mathscr{E}'(\mathbb{R}_-)$ (Yamamoto (2016)).

3. COPRIMENESS IN $\mathscr{E}'(\mathbb{R}_-)$

We first translate (2) to a divisibility condition by considering the principal ideal $(q)=q*\mathscr{E}'(\mathbb{R}_-)$ generated by q in $\mathscr{E}'(\mathbb{R}_-)$. Note first that (2) is easily seen to be equivalent to

$$p*\phi=\delta \mod q. \tag{3}$$

for some $\phi \in \mathscr{E}'(\mathbb{R}_-)$. In other words,

$$[p]*[\phi]=[\delta] \tag{4}$$

in $\mathscr{E}'(\mathbb{R}_-)/(q)$. This means that the equivalence class $[p]$ is invertible in the quotient algebra $\mathscr{E}'(\mathbb{R}_-)/(q)$.

Condition (4) by itself is not so easy to handle because of the intricate topology of $\mathscr{E}'(\mathbb{R}_-)$. However, because q has compact support, the following remarkable property holds:

Proposition 3.1. Take any $T>0$ such that $\operatorname{supp} q \subset (-T,0]$. Then

$$\mathscr{E}'(\mathbb{R}_-)/(q) \cong \mathscr{E}'([-T,0])/(q) \tag{5}$$

Proof Let π be the projection operator

$$\pi : \mathscr{D}' \to \mathscr{D}'_{(0,\infty)} : \psi \mapsto \psi|(0,\infty) \tag{6}$$

where $\mathscr{D}'_{(0,\infty)}$ is the space of distributions with support contained in $(0,\infty)$. Given $\psi \in \mathscr{D}'$, define the following operator π^q as

$$\pi^q : \mathscr{E}'(\mathbb{R}_-) \to \mathscr{E}'(\mathbb{R}_-) : \psi \mapsto q*\pi(q^{-1}*\psi). \tag{7}$$

Now for a distribution $\psi \in \mathscr{D}'_+$, define $\ell(\psi)$ as

$$\ell(\psi) := \inf\{t \in \operatorname{supp}\psi\} \tag{8}$$

where $\operatorname{supp}\psi$ denotes the support of ψ. We note from Yamamoto and Bonnet (2021) that $\ell(\alpha*\beta)=\ell(\alpha)+\ell(\beta)$ for $\alpha,\beta \in \mathscr{D}'_+$.

Take any $x \in \mathscr{E}'(\mathbb{R}_-)$ along with $\pi^q x$. We claim that $\pi^q x$ belongs to $\mathscr{E}'(\mathbb{R}_-)$ (hence (7) is well defined as a map from $\mathscr{E}'(\mathbb{R}_-)$ into itself) and that $x \cong \pi^q x \bmod q$. We have

$$q^{-1}*(x-\pi^q x)=q^{-1}*x-q^{-1}*q*\pi(q^{-1}x)=q^{-1}*x-\pi(q^{-1}*x).$$

The last term $\phi := q^{-1}*x-\pi(q^{-1}*x)$ belongs to $\mathscr{E}'(\mathbb{R}_-)$ because $q^{-1}*x-\pi q^{-1}*x$ must be zero on $(0,\infty)$. That is to say,

$$x-\pi^q x = q*\phi \in q*\mathscr{E}'(\mathbb{R}_-)=(q).$$

This also shows that $\pi^q x = x-q*\phi \in \mathscr{E}'(\mathbb{R}_-)$. In other words, $[x]=[\pi^q x]$ in $\mathscr{E}'(\mathbb{R}_-)/(q)$. Moreover, since $\ell(\pi(q^{-1}*x)) \geq 0$ and $\ell(q) \geq -T$, the support of $\pi^q x = q*\pi(q^{-1}*x)$ must be contained in $[-T,0]$ by $\ell(q*\pi(q^{-1}*x))=\ell(q)+\ell(\pi(q^{-1}*x))$. That is, for every $x \in \mathscr{E}'(\mathbb{R}_-)$, there always exists an element $\pi^q x$ such that $\operatorname{supp}\pi^q x \subset [-T,0]$, and $x \cong \pi^q x \bmod q$. This proves (5). □

Remark 3.2. Proposition 3.1 claims that as far as a pseudorational impulse response is concerned, we can confine our attention to those inputs with support contained in $[-T,0]$ with $-T<\ell(q)$. This result is not so surprising if we pay proper attention to the compact-support property of q. Since q has bounded support, its maximum length should determine the maximum length of memory needed to reconstruct the state or future outputs. This can be easily guessed once we resort to the analogy with realization theory for discrete-time linear systems: The degree of the denominator polynomial $q(z)$ determines the dimension of the state in the standard reachable realization, and the degree here exactly corresponds to the length of the support of q here. The projection scheme used above is an analogy to the finite-dimensional theory developed by Fuhrmann (1976).

4. GEL'FAND ALGEBRA STRUCTURE OF THE SPACE OF MEASURES

We have seen that the existence of the Bézout condition reduces to the invertibility of $[p]$ in the quotient algebra $\mathscr{E}'(\mathbb{R}_-)/(q)$. It is also seen that this space $\mathscr{E}'(\mathbb{R}_-)/(q)$ is isomorphic to $\mathscr{E}'([-T,0])/(q)$ for some $T>0$ so that its structure is quite simplified. However, the space $\mathscr{E}'(\mathbb{R}_-)/(q)$ is still not that easy to tackle due to a rather complex topological structure of $\mathscr{E}'(\mathbb{R}_-)/(q)$.

We now choose to confine ourselves to the subspace $\mathfrak{M}(\mathbb{R}_-)$ that is the subspace of $\mathscr{E}'(\mathbb{R}_-)$ consisting of measures, i.e., those with elements of order 0. As shown in Proposition 3.1, $\mathfrak{M}(\mathbb{R}_-)/(q) \cong \mathfrak{M}([-T,0])/(q)$ for some $T>0$. (Proposition 3.1 claims this fact for $\mathscr{E}'(\mathbb{R}_-)$, but the proof remains essentially the same.) Note that $\operatorname{ord} q^{-1} = -\operatorname{ord} q = 0$ by condition (2) of Definition 2.1, so that q^{-1} is also a measure.) We here observe that the space $\mathfrak{M}([-T,0])/(q)$ has a remarkable advantage over $\mathscr{E}'([-T,0])/(q)$ in that it can be regarded as a Banach space with respect to the strong dual topology as the dual space of the space of continuous functions $C[-T,0]$. Furthermore, it inherits a natural algebra structure induced from $\mathfrak{M}(\mathbb{R}_-)$ (with respect to convolution) with the unity element $[\delta]$. In other words, it is a Gel'fand algebra (Gel'fand et al. (1964); Berberian (1973)).

A Gel'fand algebra is known to have a remarkable property in that the invertibility of an element can be well tested by characterizing the space of its maximal ideals (Berberian (1973); Gel'fand et al. (1964)). This fact is best suited to study the invertibility condition (4).

Let us now make the following Assumption:

Assumption 1 There exists $\sigma \in \mathbb{R}$ such that $\hat{p}(s)$ and $\hat{q}(s)$ do not vanish on $\{s|\operatorname{Re} s \geq \sigma\}$.

For the validity of Assumption 1, we note the following. Since (p,q) is a pseudorational pair, $q^{-1} \in \mathscr{D}'_+$. Then there exists $\sigma \in \mathbb{R}$ such that $1/\hat{q}(s)$ is of polynomial order for $\operatorname{Re} s \geq \sigma$ according to Schwartz (1961). Hence $\hat{q}(s)$ do not vanish for $\operatorname{Re} s \geq \sigma$. Likewise, if there exists $p^{-1} \in \mathscr{D}'_+$, the same is true of $\hat{p}(s)$.

If Assumption 1 is satisfied we may assume that σ can be taken to be zero, without loss of generality. For if necessary, we can always shift the complex variable as $s \mapsto s-\sigma$, and this clearly does not affect the coprimeness relationship.

The following theorem was first given in Yamamoto (2007), but we here give a more complete proof for the sake of completeness.

Theorem 4.1. Let $p,q \in \mathfrak{M}(\mathbb{R}_-)$, and satisfy Assumption 1. Suppose that there exists $c > 0, a \in \mathbb{R}$ such that

$$|\hat{p}(s)| + |\hat{q}(s)| \geq c > 0 \tag{9}$$

for every $s \in \mathbb{C}_- = \{s \in \mathbb{C} | \operatorname{Re} s \leq 0\}$. Then the (p,q) is a Bézout pair, i.e., satisfies the Bézout identity (2).

For the proof, we need some preliminaries. The question here is to find a condition under which $[p]$ is invertible in $\mathfrak{M}([-T,0])/(q)$. By Gel'fand representation theory (Berberian (1973); Gel'fand et al. (1964)), an element $[p]$ is invertible if and only if it belongs to no maximal ideals.

Consider the Laplace transform of elements in $\mathfrak{M}(\mathbb{R}_-)$. It is easy to see that this is a subalgebra of $H^\infty(\mathbb{C}_-)$. Then, as in Hoffman (1962), we see that the correspondence

$$\psi \mapsto \hat{\psi}(s)$$

considered for $s \in \mathbb{C}_-$ gives the Gel'fand representation.

What is then a maximal ideal in $\mathfrak{M}(\mathbb{R}_-)$? Take any $\lambda \in \mathbb{C}_-$, and consider the point evaluation

$$\phi_\lambda : f \mapsto \hat{f}(\lambda). \tag{10}$$

It is easy to see that ϕ_λ is a complex homomorphism (i.e., homomorphism from $\mathfrak{M}(\mathbb{R}_-)$ to $\mathbb{C}$), and hence $\ker \phi_\lambda$ is a maximal ideal of $\mathfrak{M}(\mathbb{R}_-)$. Observe however that this does not necessarily yield a maximal ideal in $\mathfrak{M}(\mathbb{R}_-)/(q)$, because in order to be an ideal in this space, this ideal should contain (q). In other words, $\hat{q}$ should vanish there. If M is given by

$$M_\lambda = \{f | \hat{f}(\lambda) = 0\},$$

then this means that λ should be a zero of $\hat{q}$ for $M_\lambda \supset (q)$. Now let

$$\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots \tag{11}$$

be the set of zeros of $\hat{q}$. Then we have maximal ideals

$$M_{\lambda_1}, M_{\lambda_2}, \ldots, M_{\lambda_n}, \ldots$$

of $\mathfrak{M}(\mathbb{R}_-)/(q)$. But these are not all. There are other maximal ideals that are centered at "infinity".

To see this, let us first start with the following proposition:

Proposition 4.2. Let $f \in \mathfrak{M}(\mathbb{R}_-)$, and suppose that $\phi(f) = 0$ for some complex homomorphism, i.e., f belongs to a maximal ideal $\ker \phi$. Suppose also that ϕ does not agree with any of M_{λ_n} as given above. Then there exists a sequence μ_n such that

- $\mu_n \to \infty$ and
- $\hat{f}(\mu_n) \to 0$ as $n \to \infty$

Proof Suppose there exists no such μ_n. Then there exists $\delta > 0$ and $R > 0$ such that $|\hat{f}(s)| \geq \delta$ for $|s| \geq R$. In view of the Hadamard factorization that gives rise to an infinite product representation of $\hat{q}(s)$ as linear factors of $1 - s/\mu_n$ (Boas (1954)), it follows that either

(1) $\hat{f}(s)$ has infinitely many zeros, or
(2) $\hat{f}(s)$ has only finitely many zeros.

The first case is clearly impossible by $|\hat{f}(s)| \geq \delta$. Hence $\hat{f}$ has only finitely many zeros. But this yields $\hat{f}(s) = e^{\alpha s} P(s)$ where P is a polynomial. Note that $\alpha \geq 0$ because the inverse Laplace transform of $\hat{f}$ is a measure in $\mathfrak{M}(\mathbb{R}_-)$. Since $\alpha = 0$ just corresponds to a constant, we assume $\alpha \neq 0$, so that $\alpha > 0$. But then $e^{\alpha s}$ can have infinitely many zeros along the imaginary axis, and this contradicts $|\hat{f}(s)| \geq \delta$ for $|s| \geq R$. Hence $\hat{f}$ must be a polynomial. But this is again impossible unless $\hat{f}$ is a (nonzero) constant because the inverse Laplace transform of $\hat{f}$ must be a measure. Therefore $\hat{f}$ must be a constant. But this yields $\phi(1) = 0$, which clearly means that ϕ annihilates the whole space, and this contradicts the fact that ϕ is a nontrivial complex homomorphism (or $\ker \phi$ is a maximal ideal). □

In particular, this holds also for q. Then if M is a maximal ideal of $\mathfrak{M}(\mathbb{R}_-)/(q)$, then $\pi^{-1}(M)$ is clearly a maximal ideal of $\mathfrak{M}(\mathbb{R}_-)$, and this should contain (q).

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1 Suppose (9) holds, but p belongs to a maximal ideal in $\mathfrak{M}(\mathbb{R}_-)/(q)$. If p belongs to one of M_{λ_n}, then this would clearly contradict (9). Hence assume that $\hat{p}$ vanishes at no λ_n, $n = 1, 2, \ldots$. Then by Proposition 4.2, there exists μ_n such that $\mu_n \to \infty$ and $\hat{p}(\mu_n) \to 0$. Since this maximal ideal should contain q, $\hat{q}$ should also vanish there, and hence a suitable subsequence of $\hat{q}(\mu_n)$ should go to 0. This clearly contradicts (9). □

Here are some examples:

Example 4.3. The pair $(e^{s/2} - 1, e^s - 1)$ is not a Bezout pair. The pair possesses infinitely many common zeros.

Example 4.4. The pair $(e^s, e^{s/2} - 1)$ is a Bezout pair. It is easy to check (9). This can also be directly verified by $e^s - (e^{s/2} - 1)(e^{s/2} + 1) = 1$.

Remark 4.5. Condition (9) is the same as that in the celebrated Corona theorem by Carleson for H^∞ (Duren (1970); Garnett (1981)). One should of course be careful not to confuse the present result with the Corona theorem, because such conditions crucially depend on the choice of a ring. The proof here is good deal simpler than that of the monstrous Corona theorem (Duren (1970); Garnett (1981)). This is because the algebra $\mathfrak{M}(\mathbb{R}_-)/(q)$ is much "smaller" than H^∞, and the way it yields "cancellation at infinity" is quite much restricted by the discrete zeros $\{\lambda_n\}$ whereas in the case of the Corona theorem, there are almost arbitrary ways in which such sequences go to infinity.

5. EXTENSION TO $\mathscr{E}'(\mathbb{R}_-)$

It is thus quite tempting to try to generalize the above result to the general case of $\mathscr{E}'(\mathbb{R}_-)$ or $\mathscr{E}'(\mathbb{R})$.

We first make the following assumption:

Assumption 2: The algebraic multiplicity of each zero λ_n of $\hat{q}(s)$ is globally bounded.

Observe the following Theorem 5.1 obtained in Yamamoto (2007, 2016).

Theorem 5.1. Let $q^{-1} * p$ be pseudorational, and suppose that there exists a nonnegative integer m such that

$$|\lambda_n^m \hat{p}(\lambda_n)| \geq c > 0, n = 1, 2, \ldots \tag{12}$$

Then the pair (p,q) satisfies the Bézout identity (2) for some $\phi, \psi \in \mathscr{E}'(\mathbb{R}_-)$.

The proof given in Yamamoto (2007, 2016) is fairly complicated and highly technical. It does involve some elaborate analysis of complex analytic functions of exponential type, and some deep facts of their growth orders.

It is thus tempting to try to give a proof by using Theorem 4.1, extending the result for $\mathfrak{M}(\mathbb{R}_-)$ to $\mathscr{E}'(\mathbb{R}_-)$.

Let us first prepare some pertinent facts on the structure of $\mathscr{E}'(\mathbb{R}_-)$. Since every element of $\mathscr{E}'(\mathbb{R}_-)$ has compact support,

it is of finite order (Schwartz (1966)). That is, for every $\psi \in \mathscr{E}'(\mathbb{R}_-)$, there exists $r \geq 0$ such that

$$\psi = (d/dt)^r \psi_0 \tag{13}$$

for some $\psi_0 \in \mathfrak{M}(\mathbb{R}_-)$ and $r \geq 0$. This readily implies

$$\mathscr{E}'(\mathbb{R}_-) = \cup_{r=0}^{\infty}(d/dt)\mathfrak{M}(\mathbb{R}_-). \tag{14}$$

In other words, the algebra $\mathscr{E}'(\mathbb{R}_-)$ is derived as the differentiated union of measures.

We now suppose that we are given a pseudorational pair (p,q) belonging to $\mathscr{E}'(\mathbb{R}_-)$. Since $\mathscr{E}'(\mathbb{R}_-)$ is the nested union of differentiated measures, we may hope that we can reduce the coprimeness problem of $\mathscr{E}'(\mathbb{R}_-)$ into that of $\mathfrak{M}(\mathbb{R}_-)$. A procedure like the Euclid division algorithm can be a hint for this.

Suppose for the moment that p is of order 0 and q is of order 1. Suppose also that $\hat{q}(s)$ has one real zero, say, λ. Then the inverse Laplace transform of $\hat{q}(s)/(s-\lambda)$ should be of order zero because the division by $s-\lambda$ should act as an integration. Therefore, both p and $L^{-1}[\hat{q}(s)/(s-\lambda)]$ should be of order zero, i.e., measure.

Then it is naturally expected that the coprimeness of (p,q) should reduce to that of $(p, L^{-1}[\hat{q}(s)/(s-\lambda)])$.

In fact, if $(p, q_0 * q_1)$ is coprime in a ring R, (p,q_1) is coprime and vice versa. So it is natural to expect that the Bézout condition of (p,q) is translated to that of $(p, L^{-1}[\hat{q}(s)/(s-\lambda)])$ where the latter belong to the space of measures $\mathfrak{M}(\mathbb{R}_-)$, where Theorem 4.1 is available.

However, this seemingly reasonable idea unfortunately does not work. The following counterexample shows why.

Example 5.2. Consider the pair $(\delta'_{-1} - \delta, \delta_{-1})$. This pair is clearly pseudorational. The element $\delta'_{-1} - \delta$ has order 1, and δ_{-1} has order 0, i.e., measure. They admit Laplace transforms $se^s - 1$ and e^s, respectively. They satisfy the Bézout identity

$$(se^s - 1)\cdot(-1) + s \cdot e^s = 1, \tag{15}$$

or

$$(\delta'_{-1} - \delta) * (-\delta) + \delta' * \delta_{-1} = \delta, \tag{16}$$

and hence the pair is coprime over $\mathscr{E}'(\mathbb{R}_-)$.

The former element $se^s - 1$ has one positive zero, say α. This means that $(se^s - 1)/(s-\alpha)$ (or its inverse Laplace transform) has order 0 because division by $s-\alpha$ entails in integration of $\delta'_{-1} - \delta$ once, whereby yielding an element of order zero, i.e., a measure.

In other words, the pair (or the respective inverse Laplace transforms) $(se^s - 1)/(s-\alpha), e^s)$ belongs to $\mathfrak{M}(\mathbb{R}_-)$, and they are coprime over $\mathscr{E}'(\mathbb{R}_-)$. However, this does not guarantee that this pair admits a coprime factorization over $\mathfrak{M}(\mathbb{R}_-)$ in the sense of Theorem 4.1.

To see this, observe that $se^s - 1$ admits infinitely many zeros λ_n such that $\mathrm{Re}\,\lambda_n \to -\infty$. (This can easily be seen by noting that it is the characteristic function of the retarded delay-differential equation $\dot{x} = x(t-1) + u$.) Indeed, $\lambda_n e^{\lambda_n} = 1$ admits infinitely many solutions such that $e^{\lambda_n} = 1/\lambda_n, n = 1, 2, \ldots$. This also implies that $\hat{p}(\lambda_n) \to 0$ as $n \to \infty$. That is, it contradicts condition (9) of Theorem 4.1, and cannot be a Bézout pair in $\mathfrak{M}(\mathbb{R}_-)$.

In other words, the pair can admit a Bézout identity over $\mathscr{E}'(\mathbb{R}_-)$ with $x, y \in \mathscr{E}'(\mathbb{R}_-)$, but it cannot satisfy a Bézout condition over the algebra of $\mathfrak{M}(\mathbb{R}_-)$ because the latter algebra is much smaller than $\mathscr{E}'(\mathbb{R}_-)$ and does not give as much freedom as that induced by $\mathscr{E}'(\mathbb{R}_-)$. This can be more directly seen by noting the identity $s \cdot e^s + (se^s - 1)/(s-\alpha) \cdot (\alpha - s) = 1$ This looks trivial and not any different from (15). The difference here is that the multiplying factor $\alpha - s$ that makes the pair $(se^s - 1)/(s-\alpha), e^s)$ satisfy the Bézout identity does not belong to (the Laplace transform of) $\mathfrak{M}(\mathbb{R}_-)$. To cover this situation, we do need Theorem 5.1, which cannot be, unfortunately, covered as a natural variant of Theorem 4.1.

In fact, at the zeros λ_n of $\hat{q}$, $e^{\lambda_n} = 1/\lambda_n$ holds, so that the $\hat{p}(\lambda_n)$ clearly satisfy condition (12) for $m = 1$. This condition can also be rewritten as

$$|s\hat{p}(s)| + |s\hat{q}(s)| \geq c > 0, \forall s \in \mathbb{C}_-. \tag{17}$$

6. CONCLUDING REMARKS

We have seen that the space $\mathfrak{M}(\mathbb{R}_-)$ of measures admits a Gel'fand algebra structure, and it yields a concrete Corona-like condition (9) for the Bézout identity for a pseudorational pair (p,q). We have also pursued to derive the general condition (12) for the Bézout identity over $\mathscr{E}'(\mathbb{R}_-)$, but also seen that a straightforward reduction idea does not work. The modified generalized Corona-like condition (17) may, however, suggest that there could still be a possibility of generalizing (9) to a more general context in $\mathscr{E}'(\mathbb{R}_-)$.

REFERENCES

Berberian, Sterling, K. (1973). *Lectures in Functional Analysis and Operator Theory*. Springer.

Boas, Ralph, P.J. (1954). *Entire Functions*. Academic Press.

Duren, P.L. (1970). *Theory of H^p Spaces*. Academic Press, San Diego.

Fuhrmann, Paul, A. (1976). Algebraic system theory: an analyst's point of view. *J. Franklin Inst.*, 301, 521–540.

Garnett, J.B. (1981). *Bounded Analytic Functions*. Academic Press, San Diego.

Gel'fand, I., M., Raikov, D., A., and Shilov, G., E. (1964). *Commutative Normed Rings*. Chelsea.

Hoffman, K. (1962). *Banach Spaces of Analytic Functions*. Prentice-Hall, Englewood Cliffs, reprinted by Dover Publications.

Schwartz, L. (1961). *Méthodes Mathématiques pour les Sciences Physiques*. Hermann.

Schwartz, L. (1966). *Théorie des Distribution*. Hermann.

Yamamoto, Y. (1988). Pseudo-rational input/output maps and their realizations: a fractional representation approach to infinite-dimensional systems. *SIAM J. Control & Optimiz*, 26, 1415–1430.

Yamamoto, Y. (1989). Reachability of a class of infinite-dimensional linear systems: an external approach with applications to general neutral systems. *SIAM J. Control & Optimiz*, 27, 217–234.

Yamamoto, Y. (2007). Coprimeness in the ring of psedorational transfer functions. In *Proc. 15th Mediterranean Conf. on Control and Automation*.

Yamamoto, Y. (2016). Behavioral controllability and coprimeness for pseudorational transfer functions. *Systems & Control Letters*, 95, 20–26.

Yamamoto, Y. and Bonnet, C. (2021). Bézout identity in pseudorational transfer functions. *IFAC-PapersOnLine*, 54-9, 353–358.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Funnel control with internal model and anti-windup for input-saturated mechatronic systems

Christoph M. Hackl *

* *Hochschule München University of Applied Sciences, Lothstr. 64, 80335 München, Germany (e-mail: christoph.hackl@hm.edu).*

Abstract: Funnel control (FC) in combination with internal model (IM) achieves asymptotic tracking but feasibility of IM-FC in presence of input saturation (if e.g. a feasibility condition is satisfied) is unclear. Here, both aspects are brought together to obtain a closed-loop system comprising of funnel controller, serial interconnection of internal model with anti-windup and input-saturated high-gain stabilizable system which achieves prescribed transient and asymptotic accuracy if a feasibility conditions is satisfied (Hackl, 2017, Chapter 10). It will be illustrated that in presence of actuator saturation, funnel control with internal model but without anti-windup might exhibit integrator windup deteriorating control performance and resulting in instability of the closed-loop system. The theory of funnel control of input-saturated systems will be extended to allow for the application of funnel control with internal model to input-saturated systems by introducing an anti-windup strategy called conditional integration. The proposed approach is implemented and illustrated for a simple relative-degree-two system.

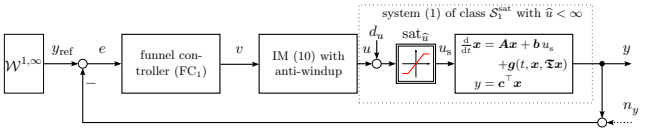
Keywords: adaptive control, funnel control, input saturation, anti-windup, internal model, feasibility condition

1. INTRODUCTION

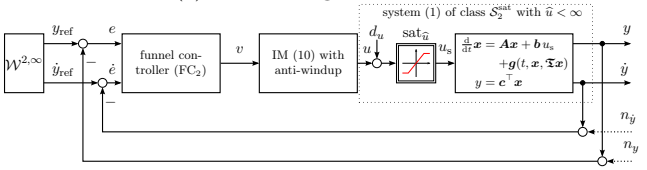
The use of funnel control (see Ilchmann et al. (2002)) in combination with internal model emerges from the well known fact that e.g. proportional-integral (PI) controllers or internal models in general are very beneficial to improve the control performance and, in particular, the *asymptotic accuracy* of closed-loop systems (see Wonham (1985); Isidori and Byrnes (1990)). That is why PI controllers are so popular in industry (see (Schröder, 2009, p. 81-82)). For speed control of mechatronic systems, this idea of connecting a PI-like internal model in series to a non-identifier based adaptive controller was first published in Schuster et al. (2004). For position control, a similar result has been published in Hackl (2011). Funnel control in conjunction with a linear internal model, i.e. *IM-funnel control*, applied to linear systems, can achieve asymptotic tracking and prescribed transient accuracy (see Ilchmann and Ryan (2006)). For mechatronic systems, funnel control with PI-like internal model, i.e. *PI-funnel control* gives steady state accuracy as well. However, it is not proven that steady state will be reached (see e.g. Ilchmann and Schuster (2009); Hackl et al. (2011); Hackl and Kennel (2012)). Except the publications Hackl (2013, 2015) on PI-funnel control *with anti-windup*, this far, from a theoretical point of view, funnel control with internal model is solely admissible for mechatronic systems *without* input saturation. In this extended abstract, it will be shown that the advantageous effects of IM-FC remain even in presence of actuator saturation if a simple conditional integration anti-windup strategy is implemented; as without it, the internal model might lead to instability of the input-saturated closed-loop system.

2. FUNNEL CONTROL WITH INTERNAL MODEL & ANTI-WINDUP FOR *INPUT-SATURATED* SYSTEMS

Funnel control is a *proportional* (and, for higher relative degrees, a proportional-derivative) control strategy. However, no integral control action is incorporated. It is well known (see Khalil (2000)) that already for exogenous signals which asymptotically converge to constant limits (e.g. *constant* references and/or disturbances), simple proportional controllers do *not* achieve steady state accuracy, i.e. $\lim_{t\to\infty} e(t) \neq 0$; the tracking error $e(t) := y_{\mathrm{ref}}(t) - y(t)$ does not tend to zero. To achieve that, at least, integral control action is required. For more complex



(a) Relative-degree-one case.



(b) Relative-degree-two case.

Fig. 1. Funnel control *with* internal model and anti-windup of *input-saturated* systems of form (1).

references and/or disturbances (e.g. for sinusoidal or ramp-like signals), internal models are beneficial and can be used in combination with funnel control. Howoever, for input-saturated systems, the interconnection of internal model

and system does *not* work as easily as for unsaturated systems (see (Hackl, 2017, Chapter 7 & 10)). The input saturation is located between internal model and input-saturated system (see Fig. 1) and might yield windup of (one or several of) the states of the internal model. Therefore, oscillations in the system output and/or system states will occur (see (Åström and Murray, 2008, Section 10.4)). Moreover, boundedness of the states of the internal model and/or closed-loop system stability are not guaranteed for input-saturated systems. It will be shown that the internal model affects the feasibility condition and a simple anti-windup strategy (conditional integration) must be adopted to ensure boundedness of the internal model states and closed-loop stability.

2.1 Considered system class

In mechatronics, the dominant dynamics of most systems can be modelled as relative-degree-one or relative-degree-two systems with stable internal dynamics (minimum-phase property), known sign of the high-frequency gain and exponentially bounded perturbation (see Hackl (2017)). The following system classes $\mathcal{S}_1^{\text{sat}}$ and $\mathcal{S}_2^{\text{sat}}$ are considered:

Let $n, m \in \mathbb{N}$, $h \geq 0$, $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n \times \mathbb{R}^n$ and $\boldsymbol{g}\colon [-h, \infty) \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$. A dynamical system, given by the functional differential equation

$$\left.\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{b}\,\text{sat}_{\widehat{u}}\big(u(t) + d_u(t)\big) + \boldsymbol{g}\big(t, \boldsymbol{x}(t), (\mathfrak{T}\boldsymbol{x})(t)\big) \\ y(t) &= \boldsymbol{c}^\top \boldsymbol{x}(t), \qquad \boldsymbol{x}|_{[-h,0]} = \boldsymbol{x}_0(\cdot) \in \mathcal{C}\big([-h, 0]; \mathbb{R}^n\big) \end{aligned}\right\} \tag{1}$$

with input saturation

$$\text{sat}_{\widehat{u}}\colon \mathbb{R} \to [-\widehat{u}, \widehat{u}],\ x \mapsto \text{sat}_{\widehat{u}}(x) := \begin{cases} \widehat{u} & , x \geq \widehat{u} \\ x & , -\widehat{u} < x < \widehat{u} \\ -\widehat{u} & , x \leq -\widehat{u} \end{cases},$$

saturation level $0 < \widehat{u} \leq \infty$, input disturbance $d_u\colon [-h, \infty) \to \mathbb{R}$, operator $\mathfrak{T}\colon \mathcal{C}([-h, \infty); \mathbb{R}^n) \to \mathcal{L}^\infty_{\text{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R}^m)$, control input $u\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ and regulated output $y(\cdot)$, is of Class $\mathcal{S}_1^{\text{sat}}$ or $\mathcal{S}_2^{\text{sat}}$ *iff the following system properties (sp) hold:*

(sp_1) *For $\mathcal{S}_1^{\text{sat}}$, the relative degree is one and the sign of the high-frequency gain is known, i.e.*

$$r = 1 \iff \gamma_0 := \boldsymbol{c}^\top \boldsymbol{b} \neq 0 \quad \textit{and} \quad \text{sign}(\gamma_0) \textit{ known}; \tag{2}$$

or, for $\mathcal{S}_2^{\text{sat}}$, the relative degree is two and the sign of the high-frequency gain is known, i.e.

$$\begin{aligned} r = 2 \iff \boldsymbol{c}^\top \boldsymbol{b} = 0 \ \wedge\ \gamma_0 := \boldsymbol{c}^\top \boldsymbol{A}\boldsymbol{b} \neq 0 \\ \wedge\ \forall (t, \boldsymbol{x}, \boldsymbol{w}) \in [-h, \infty) \times \mathbb{R}^n \times \mathbb{R}^m\colon \\ \boldsymbol{c}^\top \boldsymbol{g}(t, \boldsymbol{x}, \boldsymbol{w}) = 0 \quad \textit{and} \quad \text{sign}(\gamma_0) \textit{ known}; \end{aligned} \tag{3}$$

(sp_2) *the unperturbed system is minimum-phase, i.e.*

$$\forall\, s \in \mathbb{C}_{\geq 0}\colon \qquad \det \begin{bmatrix} s\boldsymbol{I}_n - \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{c}^\top & 0 \end{bmatrix} \neq 0; \tag{4}$$

(sp_3) *the operator is of class $\mathcal{T}$ (see Ilchmann et al. (2002)) and the input disturbance is bounded, i.e.*

$$\mathfrak{T} \in \mathcal{T} \ \textit{ and } \ d_u(\cdot) \in \mathcal{L}^\infty([-h, \infty); \mathbb{R}); \tag{5}$$

(sp_4) *the function $\boldsymbol{g}\colon [-h, \infty) \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a Caratheodory function (see Ilchmann et al. (2002)) and exponentially bounded with respect to the output $y = \boldsymbol{c}^\top \boldsymbol{x}$, i.e. for unknown $q \geq 0$, the following holds*

$$\begin{aligned} \exists M_{\boldsymbol{g}} > 0\, \exists q \geq 0 \,\text{for a.a. } t \in [-h, \infty)\, \forall (\boldsymbol{x}, \boldsymbol{w}) \in \mathbb{R}^n \times \mathbb{R}^m\colon \\ \|\boldsymbol{g}(t, \boldsymbol{x}, \boldsymbol{w})\| \leq M_{\boldsymbol{g}}\big[1 + \exp\big(|\boldsymbol{c}^\top \boldsymbol{x}|^q\big)\big]; \end{aligned} \tag{6}$$

(sp_5) *for $\mathcal{S}_1^{\text{sat}}$, the regulated output $y(\cdot)$ is available for feedback; or, for $\mathcal{S}_2^{\text{sat}}$, the regulated output $y(\cdot)$ and its derivative $\dot{y}(\cdot)$ are available for feedback*

System (1) can represent nonlinear input-saturated dynamical mechatronic systems with bounded input disturbance and exponentially bounded nonlinear perturbations. System examples are speed, position or current controlled electrical drives or elastic servo systems (see Hackl (2017)). Due to the nonlinear perturbation function $\boldsymbol{g}(\cdot, \cdot, \cdot)$, the system dynamics are nonlinear in exogenous (time-varying) signals, system state $\boldsymbol{x}(\cdot)$ and functional perturbation $(\mathfrak{T}\boldsymbol{x})(\cdot)$. Input disturbance $d_u(\cdot)$ in (1) allows to incorporate bounded actuator deviations and/or feedforward commands, whereas time-dependent and functional perturbations in (1) account for e.g. time-varying electrical/mechanical loads and nonlinear friction effects.

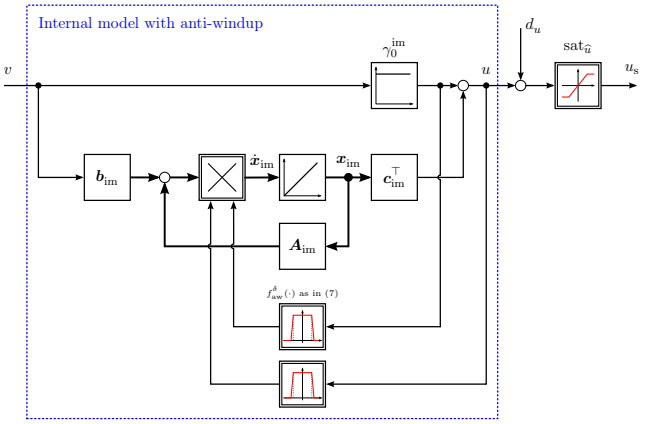


Fig. 2. Internal model (10) with anti-windup due to decision functions (7).

2.2 Funnel controllers for systems of class $\mathcal{S}_1^{\text{sat}}$ and $\mathcal{S}_2^{\text{sat}}$

Despite the existence of several funnel controller variants, in this paper, the most simple funnel controllers with gain scaling for systems of class $\mathcal{S}_1^{\text{sat}}$ and $\mathcal{S}_2^{\text{sat}}$ with tracking error $e(t) = y_{\text{ref}}(t) - y(t)$ are considered:

- Funnel controller for systems of class $\mathcal{S}_1^{\text{sat}}$:

$$v(t) = \text{sign}(\gamma_0) k(t) e(t) \ \textit{ where } \ k(t) = \frac{\varsigma(t)}{\psi(t) - |e(t)|} \tag{FC$_1$}$$

- Funnel controller for systems of class $\mathcal{S}_2^{\text{sat}}$:

$$\begin{aligned} &v(t) = \text{sign}(\gamma_0)\Big(k_0(t)^2 e(t) + k_0(t) k_1(t) \dot{e}(t)\Big) \text{ where} \\ &k_0(t) = \frac{\varsigma_0(t)}{\psi_0(t) - |e(t)|} \text{ and } k_1(t) = \frac{\varsigma_1(t)}{\psi_1(t) - |\dot{e}(t)|}. \end{aligned} \tag{FC$_2$}$$

The gain scaling functions $\varsigma(\cdot)$, $\varsigma_1(\cdot)$ and $\varsigma_2(\cdot)$ and the funnel boundaries $\psi(\cdot)$, $\psi_0(\cdot)$ and $\psi_1(\cdot)$ are element of the Sobolov space (i.e. $\mathcal{W}^{1,\infty}(\mathbb{R}_{\geq 0}, [\lambda, \infty))$), bounded away from zero and have essentially bounded derivatives (for details see (Hackl, 2017, Chapter 9)). The reference signals $y_{\text{ref}}(\cdot)$ are element of $\mathcal{W}^{1,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R})$ for $\mathcal{S}_1^{\text{sat}}$ systems and of $\mathcal{W}^{2,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R})$ for $\mathcal{S}_2^{\text{sat}}$ systems (see Ilchmann et al. (2002) and Hackl et al. (2013), respectively).

2.3 Internal model with anti-windup

In order to guarantee anti-windup and boundedness of the states of the internal model, a simple conditional integration approach is proposed. To do so, the internal model must be equipped with two anti-windup decision functions (see Fig. 2). The proposed *anti-windup decision functions* are Lipschitz continuous. i.e. $f^{\delta(\cdot)}_{\Delta,\widehat{u}}\colon \mathbb{R}\to[0,1]$,

$$u \mapsto f^{\delta(\cdot)}_{\Delta,\widehat{u}}(u) := \begin{cases} 0, & u < -\widehat{u} \\ \delta(u), & -\widehat{u} \le u \le -\widehat{u}+\Delta \\ 1, & -\widehat{u}+\Delta < u < \widehat{u}-\Delta \\ \delta(-u), & \widehat{u}-\Delta \le u \le \widehat{u} \\ 0, & u > \widehat{u}, \end{cases} \tag{7}$$

where $\Delta > \widehat{u}$ and, for $\mathbb{I}_\Delta := [-\widehat{u}, -\widehat{u}+\Delta] \cup [\widehat{u}-\Delta, \widehat{u}]$,

$$\delta(\cdot) \in \left\{ f(\cdot) \in \mathcal{C}^{\mathrm{L}}\big(\mathbb{I}_\Delta;[0,1]\big) \;\middle|\; \begin{array}{l} f(-\widehat{u}) = f(\widehat{u}) = 0, \;\text{ and} \\ f(\Delta-\widehat{u}) = f(\widehat{u}-\Delta) = 1 \end{array} \right\}.$$

For examples, please refer to (Hackl, 2017, Chapter 10). The following lemma shows that for a proper internal model design in combination with two decision functions, the output of the internal will remain bounded for all time and for any continuous input.

Lemma 1. (Internal model with anti-windup). For $p \in \mathbb{N}$,

$$\left.\begin{aligned} \boldsymbol{A}_{\mathrm{im}} &:= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ -\widehat{a}_0 & -\widehat{a}_1 & \dots & -\widehat{a}_{p-2} & -\widehat{a}_{p-1} \end{bmatrix} \in \mathbb{R}^{p\times p}, \\ \boldsymbol{b}_{\mathrm{im}} &:= \begin{pmatrix} 0, & \cdot, & \cdot, & 0, & 1 \end{pmatrix}^\top \in \mathbb{R}^p, \;\text{ and} \\ \boldsymbol{c}_{\mathrm{im}} &:= \begin{pmatrix} \widehat{c}_0, & \dots, & \widehat{c}_{p-1} \end{pmatrix}^\top \in \mathbb{R}^p, \end{aligned}\right\} \tag{8}$$

let the internal model

$$\left.\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_{\mathrm{im}}(t) &= \boldsymbol{A}_{\mathrm{im}}\boldsymbol{x}_{\mathrm{im}}(t) + \boldsymbol{b}_{\mathrm{im}}\, v(t) \\ u(t) &= \boldsymbol{c}_{\mathrm{im}}^\top \boldsymbol{x}_{\mathrm{im}}(t) + \gamma_0^{\mathrm{im}} v(t) \end{aligned}\right| \begin{array}{l} \deg(D_{\mathrm{im}}) =: p \in \mathbb{N}, \\ \boldsymbol{x}_{\mathrm{im}}(0) = \boldsymbol{x}^0_{\mathrm{im}} \in \mathbb{R}^p, \\ \operatorname{sign}(\gamma_0^{\mathrm{im}}) = \operatorname{sign}(\gamma_0) \end{array} \tag{9}$$

be a minimal realization. With *anti-windup*, it is given by

$$\left.\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_{\mathrm{im}}(t) &= f^{\delta(\cdot)}_{\Delta,\widehat{u}}\Big(\boldsymbol{c}_{\mathrm{im}}^\top \boldsymbol{x}_{\mathrm{im}}(t) + \gamma_0^{\mathrm{im}} v(t)\Big) \cdot f^{\delta(\cdot)}_{\Delta,\widehat{u}}\Big(\gamma_0^{\mathrm{im}} v(t)\Big) \cdot \\ &\quad \cdot \Big[\boldsymbol{A}_{\mathrm{im}}\boldsymbol{x}_{\mathrm{im}}(t) + \boldsymbol{b}_{\mathrm{im}}\, v(t)\Big] \\ u(t) &= \underbrace{\boldsymbol{c}_{\mathrm{im}}^\top \boldsymbol{x}_{\mathrm{im}}(t)}_{=:u_{\mathrm{im}}(t)} + \gamma_0^{\mathrm{im}} v(t), \;\; \boldsymbol{x}_{\mathrm{im}}(0) = \boldsymbol{x}^0_{\mathrm{im}} \in \mathbb{R}^p \end{aligned}\right| \text{with } (\boldsymbol{A}_{\mathrm{im}}, \boldsymbol{b}_{\mathrm{im}}, \boldsymbol{c}_{\mathrm{im}}) \text{ as in (8) and } f^{\delta(\cdot)}_{\Delta,\widehat{u}}(\cdot) \text{ as in (7)} \tag{10}$$

and, for $\widehat{u} > \Delta > 0$, $v(\cdot) \in \mathcal{C}(\mathbb{R}_{\ge 0};\mathbb{R})$, the following hold:

(i) there exists a unique solution $\boldsymbol{x}_{\mathrm{im}}\colon [0,T) \to \mathbb{R}^p$, $T \in (0,\infty]$ which can be maximally extended;
(ii) the solution $\boldsymbol{x}_{\mathrm{im}}\colon [0,T) \to \mathbb{R}^p$ is global, i.e. $T = \infty$;
(iii) the sub-output $u_{\mathrm{im}}(\cdot) = \boldsymbol{c}_{\mathrm{im}}^\top \boldsymbol{x}_{\mathrm{im}}(\cdot)$ of the internal model (10) is uniformly bounded, i.e.

$$\forall\, t \ge 0\colon\; |u_{\mathrm{im}}(t)| \le M_{u_{\mathrm{im}}} := \max\{\widehat{u}, |\boldsymbol{c}_{\mathrm{im}}^\top \boldsymbol{x}^0_{\mathrm{im}}|\} + \widehat{u}; \tag{11}$$

(iv) there exists $M_{\boldsymbol{x}_{\mathrm{im}}} \ge 1$ such that $\|\boldsymbol{x}_{\mathrm{im}}(t)\| \le M_{\boldsymbol{x}_{\mathrm{im}}}$ for all $t \ge 0$, if the polynomial $\widehat{N}_{\mathrm{im}}(s) = \widehat{c}_{p-1}\, s^{p-1} + \dots + \widehat{c}_1\, s + \widehat{c}_0$ is Hurwitz.

The proof can be found in (Hackl, 2017, Lemma 10.6).

2.4 Closed-loop system

In Hopfe et al. (2010) and Hackl et al. (2013), it was shown that funnel control is applicable for input-saturated relative-degree-one and relative-degree-two systems (of e.g. class $\mathcal{S}_1^{\mathrm{sat}}$ or $\mathcal{S}_2^{\mathrm{sat}}$), if a feasibility condition $\boxed{\widehat{u}_{\mathrm{feas}} \le \widehat{u} < \infty}$ is satisfied. The feasibility bound $\widehat{u}_{\mathrm{feas}}$ depends on system parameters, disturbances, perturbations, internal dynamics and controller design (e.g. funnel boundary). It is usually (very) conservative. In view of Lemma 1, the internal model with anti-windup is uniformly bounded and can be considered as globally bounded input perturbation to the input-saturated system of class $\mathcal{S}_1^{\mathrm{sat}}$ or $\mathcal{S}_2^{\mathrm{sat}}$. Therefore, the feasibility bound $\widehat{u}_{\mathrm{feas}}$ must be adjusted and will also depend on the output bound $M_{u_{\mathrm{im}}}$ as in (11) of the internal model. Due to space limitations, full theorems and proofs are omitted but can be found in Chapter 10 of Hackl et al. (2017) for system class $\mathcal{S}_1^{\mathrm{sat}}$ and $\mathcal{S}_2^{\mathrm{sat}}$ (see Theorem 10.7 and Theorem 10.9, respectively).

3. IMPLEMENTATION AND SIMULATION RESULTS

To illustrate the difference between the closed-loop performance of IM-funnel controller (FC$_2$)+(9) *without* anti-windup [——] and IM-funnel controller (FC$_2$)+(10) *with* anti-windup [——], both are applied to a simple system

$$\left.\begin{aligned} \ddot{y}(t) &= \gamma_0 \underbrace{\operatorname{sat}_{\widehat{u}}\big(u(t)\big)}_{=:u_{\mathrm{s}}(t)}, \\ &\text{with } \gamma_0 = 3 \text{ and } \widehat{u} = 7, \end{aligned}\right| \; (y(0),\dot{y}(0))^\top = (0,0)^\top \in \mathbb{R}^2, \tag{12}$$

with relative degree two and input saturation. The saturated input is denoted by u_{s}. Output $y(\cdot)$ and its derivative $\dot{y}(\cdot)$ are available for feedback. It is easy to see that, for known $\operatorname{sign}(\gamma_0)$, system (12) is element of class $\mathcal{S}_2^{\mathrm{sat}}$ and, hence, IM-funnel control (FC$_2$)+(10) *with* anti-windup is admissible if $\widehat{u}$ is sufficiently large. The chosen reference $y_{\mathrm{ref}}(\cdot)$ (see Fig. 3a) allows for an internal model design to reduplicate a constant signal and a sinusoidal signal with angular frequency $\omega_0 = 2\pi 0.5 = \pi\,\frac{\mathrm{rad}}{\mathrm{s}}$. The design yields

$$F_{\mathrm{im}}(s) = 1 + \frac{9\,s^2 + (27-\omega_0^2)\,s + 27}{s^3 + \omega_0^2 s} =: 1 + \frac{\widehat{N}_{\mathrm{im}}(s)}{D_{\mathrm{im}}(s)} \tag{13}$$

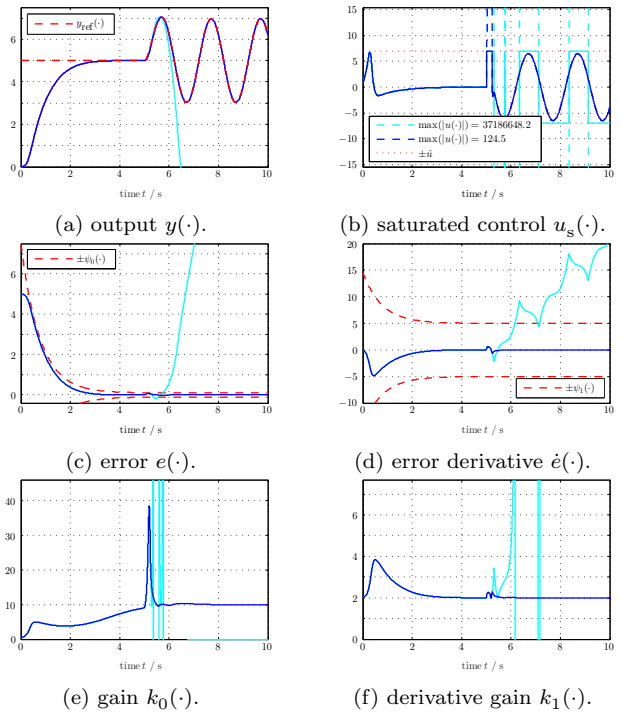
in the frequency domain. Clearly, (13) has positive high-frequency gain $\gamma_0^{\mathrm{im}} = 1$ and it is minimum-phase as the numerator $\widehat{N}_{\mathrm{im}}(s) = 9\,s^2 + (27-\omega_0^2)\,s + 27$ is Hurwitz for all $\omega_0^2 < 27$. To implement the internal model (10) with anti-windup in state space, a minimal realization of (13) must be found resulting in the following internal model system matrix and input and output coupling vectors

$$\boldsymbol{A}_{\mathrm{im}} := \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -\omega_0^2 & 0 \end{bmatrix},\; \boldsymbol{b}_{\mathrm{im}} := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \text{ and } \boldsymbol{c}_{\mathrm{im}} := \begin{pmatrix} 27 \\ (27-\omega_0^2) \\ 9 \end{pmatrix}. \tag{14}$$

The implemented anti-windup decision function $f^{\delta(\cdot)}_{\Delta,\widehat{u}}(\cdot)$ has the parameters $\Delta = 0.5$ and $\widehat{u} = 7$ and the function

$$\delta\colon \mathbb{I}_\Delta \to [0,1],\; u \mapsto \delta(u) := \tfrac{1}{2}\Big(\sin\big(\tfrac{\pi}{\Delta}(u+\widehat{u}) - \tfrac{\pi}{2}\big) + 1\Big)$$

was used. The funnel controller (FC$_2$) is equipped with exponential funnel boundary $(\psi_0(t), \psi_1(t)) := \big((\Lambda_0 - \lambda_0)\exp\left(-\frac{t}{T_{\mathrm{exp}}}\right) + \lambda_0,\, \frac{\Lambda_0-\lambda_0}{T_{\mathrm{exp}}}\exp\left(-\frac{t}{T_{\mathrm{exp}}}\right) + \lambda_1\big)^\top$ (with $\Lambda_0 = 7.5$, $\lambda_0 = 0.1$, $T_{\mathrm{exp}} = 0.77\,\mathrm{s}$ and $\lambda_1 = 5$) and gain scaling functions $\varsigma_0(t) = \psi_0(t)$ and $\varsigma_1(t) = 2\,\psi_1(t) =$

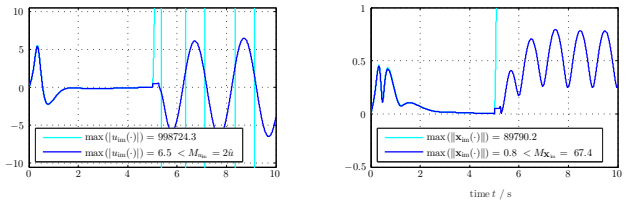


(a) output $y(\cdot)$. (b) saturated control $u_{\mathrm{s}}(\cdot)$.

(c) error $e(\cdot)$. (d) error derivative $\dot{e}(\cdot)$.

(e) gain $k_0(\cdot)$. (f) derivative gain $k_1(\cdot)$.

Fig. 3. Simulation results for set-point tracking of input-saturated closed-loop systems (12), (FC_2)+(9) *without* anti-windup [——] and (12), (FC_2)+(10) *with* anti-windup [——].



(a) IM sub-output $u_{\mathrm{im}}(\cdot)$. (b) norm of IM state $\|\boldsymbol{x}_{\mathrm{im}}(\cdot)\|$.

Fig. 4. Simulation results for internal model (IM) sub-output and state of input-saturated closed-loop systems (12), (FC_2)+(9) *without* anti-windup [——] and ——(12), (FC_2)+(10) *with* anti-windup [——].

$2\psi_1(t)$. The closed-loop systems (12), (FC_2)+(9) [——] and (12), (FC_2)+(10) [——] are implemented in Matlab/Simulink with numerical solver `ode4 (Runge-Kutta)` and fixed-step size of $1\cdot10^{-4}$ s. The comparative simulation is run for 10 s. Control objective is reference tracking of $y_{\mathrm{ref}}(\cdot)$ as depicted in Fig. 3a. The simulation results are shown in Fig. 3 and Fig. 4. Due to *windup* of the internal model state $\boldsymbol{x}_{\mathrm{im}}(\cdot)$, the IM-funnel controller (FC_2)+(9) *without* anti-windup [——] becomes unstable. Its control action $u(\cdot)$ is saturated for almost all time $t \geq 5$ s. The error and its derivative cross their respective funnel boundaries at ≈ 5.8 s and ≈ 6.2 s, respectively. Accordingly, the gains $k_0(\cdot)$ and $k_1(\cdot)$ of the IM-funnel controller (FC_2)+(9) *without* anti-windup [——] change their signs and eventually diverge. In contrast, the IM-funnel controller (FC_2)+(10) *with* anti-windup [——] ensures tracking with prescribed transient accuracy. Its gains remain bounded. Moreover, due to conditional integration, the norm of the state $\|\boldsymbol{x}_{\mathrm{im}}(\cdot)\|$ and the output $u_{\mathrm{im}}(\cdot)$ of the internal model (10) *with* anti-windup [——] remains much smaller than that of IM-funnel controller (FC_2)+(9) *without* anti-windup [——]. Both bounds, $M_{u_{\mathrm{im}}} = 2\widehat{u} = 14$ and $M_{\boldsymbol{x}_{\mathrm{im}}} = 67.4$ are by far exceeded by the IM-funnel control (FC_2)+(9) *without* anti-windup [——], whereas the closed-loop system with IM-funnel controller (FC_2)+(10) *with* anti-windup [——] does *not* get close to both bounds.

4. CONCLUSION

Funnel control with internal model and anti-windup for input-saturated systems with relative degree one or two has been discussed. It has been shown that internal models are beneficial to improve asymptotic accuracy of the closed-loop system. However, in presence of input saturation the internal model must be equipped with an anti-windup strategy (conditional integration) to assure uniform boundedness of the internal model and to avoid instability of the closed-loop system. The well-known feasibility condition of funnel control for input-saturated systems must be adjusted in view of the ultimate boundedness of the internal model. Then, available results (theorems) can be re-applied to establish closed-loop system stability. Simulation results were presented to illustrate the beneficial aspects of internal models with anti-windup. The talk will discuss several mechatronic applications in more detail.

REFERENCES

Åström, K.J. and Murray, R.M. (2008). *Feedback Systems — An Introduction for Scientists and Engineers*. Princeton University Press, Princeton and Oxford.september 2012) edition.

Hackl, C.M., Kullick, J., Eldeeb, H., and Horlbeck, L. (2017). Analytical computation of the optimal reference currents for MTPC/MTPA, MTPV and MTPF operation of anisotropic synchronous machines considering stator resistance and mutual inductance. In *2017 19th European Conference on Power Electronics and Applications (EPE'17 ECCE Europe)*, P.1–P.10. doi: 10.23919/EPE17ECCEEurope.2017.8099040.

Hackl, C.M. (2011). High-gain adaptive position control. *International Journal of Control*, 84(10), 1695–1716.

Hackl, C.M. (2013). PI-funnel control with Anti-windup and its application for speed control of electrical drives. In *Proceedings of the 52nd IEEE Conference on Decision and Control (CDC)*, 6250–6255. Florence, Italy. doi: 10.1109/CDC.2013.6760877.

Hackl, C.M. (2015). Current PI-funnel control with anti-windup for synchronous machines. In *Proceedings of the 54th IEEE Conference on Decision and Control (CDC)*, 1997–2004. Osaka, Japan. doi: 10.1109/CDC.2015.7402500.

Hackl, C.M. (2017). *Non-identifier based adaptive control in mechatronics: Theory and Application*. Springer International Publishing, Berlin. doi:10.1007/978-3-319-55036-7.

Hackl, C.M., Hofmann, A.G., and Kennel, R.M. (2011). Funnel control in mechatronics: An overview. In *Proceedings of the 50th IEEE Conference on Decision and Control (CDC) and European Control Conference (ECC)*, 8000–8007. Orlando, FL, USA. doi:10.1109/CDC.2011.6160184.

Hackl, C.M., Hopfe, N., Ilchmann, A., Mueller, M., and Trenn, S. (2013). Funnel control for systems with relative degree two. *SIAM Journal on Control and Optimization*, 51(2), 965–995.

Hackl, C.M. and Kennel, R.M. (2012). Position funnel control with linear internal model. In *Proceedings of 2012 IEEE International Conference on Control Applications (CCA)*, 1334–1339. Dubrovnik, Croatia. doi: 10.1109/CCA.2012.6402679.

Hopfe, N., Ilchmann, A., and Ryan, E.P. (2010). Funnel control with saturation: Nonlinear SISO systems. *IEEE Transactions on Automatic Control*, 55(9), 2177–2182.

Ilchmann, A., Ryan, E.P., and Sangwin, C.J. (2002). Tracking with prescribed transient behaviour. *ESAIM: Control, Optimisation and Calculus of Variations*, 7, 471–493.

Ilchmann, A. and Ryan, E.P. (2006). Asymptotic tracking with prescribed transient behaviour for linear systems. *International Journal of Control*, 79(8), 910–917.

Ilchmann, A. and Schuster, H. (2009). PI-funnel control for two mass systems. *IEEE Transactions on Automatic Control*, 54(4), 918–923.

Isidori, A. and Byrnes, C.I. (1990). Output regulation of nonlinear systems. *IEEE Transactions on Automatic Control*, 35(2), 131–140.

Khalil, H. (2000). Universal integral controllers for minimum–phase nonlinear systems. *IEEE Transactions on Automatic Control*, 45(3), 490–494.

Schröder, D. (2009). *Elektrische Antriebe - Regelung von Antriebssystemen (3., bearb. Auflage)*. Springer-Verlag, Berlin.

Schuster, H., Westermaier, C., and Schröder, D. (2004). High-gain control of systems with arbitrary relative degree: Speed control for a two mass flexible servo system. In *Proceedings of the 8th IEEE International Conference on Intelligent Engineering Systems*, 486–491. Cluj-Napoca, Romania.

Wonham, W.M. (1985). *Linear Multivariable Control: A Geometric Approach*. Number 10 in Applications of Mathematics. Springer-Verlag, Berlin. 3rd edition.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Input-constrained funnel control of nonlinear systems ⋆

Thomas Berger *

* *Institut für Mathematik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany (e-mail: thomas.berger@math.upb.de).*

Abstract: We consider tracking control for uncertain nonlinear multi-input, multi-output systems modelled by functional differential equations, in the presence of input constraints. The objective is to guarantee the evolution of the tracking error within a performance funnel with prescribed asymptotic shape. We design a novel funnel controller which, in order to satisfy the input constraints, contains a dynamic component which widens the funnel boundary whenever the input saturation is active. This design is model-free, of low-complexity and extends earlier funnel control approaches.

Keywords: nonlinear systems, adaptive control, funnel control, input constraints, functional differential equations
AMS subject classifications: 93B52, 93C10, 93C40

1. INTRODUCTION

We study funnel control for the class of nonlinear systems modelled by the r-th order functional differential equation

$$\begin{aligned} y^{(r)}(t) &= f\big(d(t), T(y, \dot{y}, \ldots, y^{(r-1)})(t), u(t)\big), \\ y|_{[-h,0]} &= y^0 \in C^{r-1}([-h,0], \mathbb{R}^m), \end{aligned} \tag{1}$$

with unknown nonlinear function $f \in C(\mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^m, \mathbb{R}^m)$ and unknown operator T which satisfy a sector bound property (see Section 1.1), unknown bounded disturbance d and unknown initial trajectory y^0 in the presence of input constraints

$$u(t) = \mathrm{sat}(v(t)) \tag{2}$$

with *known* saturation function sat and control function v provided by the to-be-designed controller. Here, we propose a novel control design, which is feasible for the aforementioned class of systems, i.e., it satisfies the input constraints imposed by (2), and achieves tracking of a given reference signal with prescribed performance of the tracking error whenever the saturation is not active, that is $u(t) = v(t)$ – in this case the controller exhibits the same performance as the funnel controllers proposed in Berger et al. (2021, 2018). When the saturation is active the performance funnel, which defines the domain for the evolution of the tracking error, is widened according to a dynamic equation describing the funnel boundaries, so that the input constraints are still met. As soon as the saturation becomes inactive again, the performance funnel recovers its desired shape exponentially fast.

The concept of funnel control was developed in the seminal work (Ilchmann et al., 2002) (see also the recent survey in Berger et al. (2021)) and proved advantageous in a variety of applications such as control of industrial servo-systems (Hackl, 2017), electrical circuits (Berger and Reis, 2014), peak inspiratory pressure (Pomprapa et al., 2015) and adaptive cruise control (Berger and Rauert, 2020).

Funnel control with input saturation was first investigated in Ilchmann and Trenn (2004) for the specific application of chemical reactor models and in a more general context in Hopfe et al. (2010a,b) for systems with relative degree one and in Hackl et al. (2013) for systems with relative degree two; this approach has been applied to funnel control with anti-windup for synchronous machines in Hackl (2015). However, in the aforementioned works it was simply shown that classical funnel control is feasible for a sufficiently large saturation bound – here this bound can be arbitrarily small. Another approach to funnel control with guaranteed input constraints is bang-bang funnel control, introduced in Liberzon and Trenn (2013) for (undisturbed) nonlinear single-input, single-output systems with arbitrary relative degree. However, the bang-bang funnel control design requires various complicated feasibility assumptions and in particular the two control values must be sufficiently large.

A relative of funnel control is prescribed performance control, developed in Bechlioulis and Rovithakis (2008), see also the important work Bechlioulis and Rovithakis (2014) where the complexity issue of this approach has been solved. The problem of input constraints has been addressed within this approach e.g. in Li and Xiang (2018), where neural networks are used to approximate the nonlinearities, and in Cheng et al. (2019), where additionally a neural observer is incorporated in the controller design. In the work Wang et al. (2019) no approximations are needed (and hence the controller is of low complexity), however the proof contains an error and simulations also show that the proposed controller is infeasible in general. The problem is that the scaling parameter κ in the χ-dynamics is chosen as a constant, but actually it needs to depend on the input.

⋆ Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 471539468.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Funnel control for systems with arbitrary relative degree was considered in Berger et al. (2021, 2018). The novel input-constrained funnel control design that we propose here extends these approaches in the following aspects:

- Compared to Berger et al. (2018) a much more general class of systems is allowed here, similar to Berger et al. (2021). However, we do not require the restrictive high-gain property of the nonlinearity f or the minimum phase property (characterized by a BIBO property of the operator T) imposed in Berger et al. (2021). On the other hand, we require a sector bound property of f and T. This condition cannot be dispensed in general, because of the input saturation.
- The new controller is able to handle arbitrary input constraints (2). Even if the saturation is never active, i.e., $u(t) = v(t)$ for all $t \geq 0$ for any solution of the closed-loop system, then the new controller is able to guarantee a prescribed performance of the tracking error as in Berger et al. (2021, 2018), with exponentially decaying funnel boundaries.

1.1 System Class

We consider functional differential equations of the form (1) incorporating an operator T of the following class.

Definition 1. For $n, q \in \mathbb{N}$ and $h \geq 0$ the set $\mathbb{T}_h^{n,q}$ denotes the class of operators $T\colon C([-h,\infty),\mathbb{R}^n) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_{\geq 0},\mathbb{R}^q)$ with the following properties.

(P1) T is causal, i.e., for all ζ, $\xi \in C([-h,\infty),\mathbb{R}^n)$ and all $t \geq 0$,

$$\zeta|_{[-h,t]} = \xi|_{[-h,t]} \quad \Longrightarrow \quad T(\zeta)|_{[0,t]} = T(\xi)|_{[0,t]}.$$

(P2) T is locally Lipschitz, i.e., for each $t \geq 0$ and all $\xi \in C([-h,t],\mathbb{R}^n)$, there exist positive constants $c_0, \delta, \tau > 0$ such that, for all $\zeta_1, \zeta_2 \in C([-h,\infty),\mathbb{R}^n)$ with $\zeta_i|_{[-h,t]} = \xi$ and $\|\zeta_i(s) - \xi(t)\| < \delta$ for all $s \in [t, t+\tau]$ and $i = 1, 2$, we have

$$\operatorname*{ess\,sup}_{s\in[t,t+\tau]} \|T(\zeta_1)(s) - T(\zeta_2)(s)\| \leq c_0 \sup_{s\in[t,t+\tau]} \|\zeta_1(s) - \zeta_2(s)\|.$$

(P3) T locally maps bounded functions to bounded functions, i.e., for all $\tau > 0$ and all $c_1 > 0$, there exists $c_2 > 0$ such that, for all $\zeta \in C([-h,\tau],\mathbb{R}^n)$,

$$\sup_{t\in[-h,\tau]} \|\zeta(t)\| \leq c_1 \quad \Longrightarrow \quad \operatorname*{ess\,sup}_{t\in[0,\tau]} \|T(\zeta)(t)\| \leq c_2.$$

We stress that property (P3) in the operator class $\mathbb{T}_h^{n,q}$ is weaker than the respective property required in Berger et al. (2021, 2018), where it essentially needs to hold for "$\tau = \infty$" (and hence corresponds to a minimum phase property), while for our purposes a local version suffices.

Next we introduce a sector bound property of $f \in C(\mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^m, \mathbb{R}^m)$ and $T \in \mathbb{T}_h^{rm,q}$ as follows.

(P4) For all $y^0 \in C^{r-1}([-h,0],\mathbb{R}^m)$ there exist $M_1, \ldots, M_{r+1} \in C(\mathbb{R}_{\geq 0} \times \mathbb{R}^p \times \mathbb{R}^m, \mathbb{R}_{\geq 0})$ such that for all $t \geq 0$, all $(d, v) \in \mathbb{R}^p \times \mathbb{R}^m$ and all $\zeta_1, \ldots, \zeta_r \in C([-h,t],\mathbb{R}^m)$ we have:

$$\|f(d, T(\zeta_1, \ldots, \zeta_r)(t), v)\| \leq M_1(t,d,v) + M_2(t,d,v)\|\zeta_1|_{[-h,t]}\|_\infty + \ldots + M_{r+1}(t,d,v)\|\zeta_r|_{[-h,t]}\|_\infty$$

Note that the functions M_i in (P4) depend on the initial history y^0 in (1). We like to note that the sector bound property (P4) cannot be dispensed in the presence of (arbitrary) input constraints in general. Otherwise, already for simple system of the form

$$\dot{y}(t) = y(t)^2 + u(t), \quad y(0) = 1, \quad u(t) = \mathrm{sat}(v(t))$$

solutions may exhibit a blow-up when the input constraints are "too tight".

We are now in the position to define the class of systems to be considered here. We stress that the high-gain property of system (1) required in earlier approaches, see e.g. Berger et al. (2021), is not needed here.

Definition 2. For $m, r \in \mathbb{N}$ we say that system (1) belongs to the system class $\mathcal{N}^{m,r}$, written $(d, f, T) \in \mathcal{N}^{m,r}$, if $d \in L^\infty(\mathbb{R}_{\geq 0},\mathbb{R}^p)$, $f \in C(\mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^m, \mathbb{R}^m)$, $T \in \mathbb{T}_h^{rm,q}$ for some $p, q \in \mathbb{N}$, $h \geq 0$ and (f, T) satisfy property (P4).

In contrast to earlier approaches as in Berger et al. (2021, 2018), here we consider an additional function sat in (1), which represents an input saturation. If $\mathrm{sat} = \mathrm{id}_{\mathbb{R}^m}$, then the results from Berger et al. (2021, 2018) could be applied. For this reason, we consider a proper input saturation, which has the following, quite general, property.

(P5) $\mathrm{sat} \in C(\mathbb{R}^m, \mathbb{R}^m)$ is bounded and there exists $\theta > 0$ such that for all $v \in \mathbb{R}^m$ with $\|v\| \leq \theta$ we have $\mathrm{sat}(v) = v$.

We stress that the input saturation function sat must be known to the controller and it can be viewed as a design parameter, chosen according to the specific requirements of the application at hand. The above property (P5) allows for a large variety of possible saturations, apart from the standard saturation $\mathrm{sat}_i(v) = v_i$ for $|v_i| \leq M$ and $\mathrm{sat}_i(v) = \mathrm{sgn}(v_i)M$ for $|v_i| > M$ for all $i = 1, \ldots, m$.

1.2 Control objective

The objective is to design a dynamic output derivative feedback strategy such that for any reference signal $y_{\mathrm{ref}} \in C^r(\mathbb{R}_{\geq 0},\mathbb{R}^m)$ the tracking error $e = y - y_{\mathrm{ref}}$ evolves within a performance funnel

$$\mathcal{F}_\psi := \{\, (t,e) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^m \mid \|e\| < \psi(t) \,\},$$

see Fig. 1, which has a prescribed shape of the form $\psi(t) = ae^{-bt} + c$ whenever the saturation in (2) is not active, i.e., $\mathrm{sat}(v(t)) = v(t)$, and $\psi(t)$ is allowed to become larger when the saturation is active. The specific shape of the performance funnel should be determined by a dynamic part of the control law.

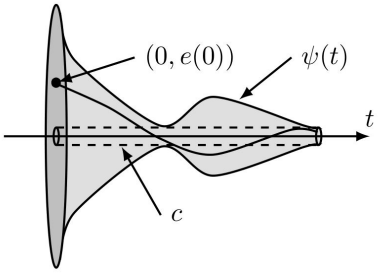


Fig. 1. Error evolution in a funnel $\mathcal{F}_\psi$ with boundary $\psi(t)$.

It is usually the hallmark of funnel control that the funnel boundary is prescribed a priori and can be freely chosen by

the designer, see e.g. Berger et al. (2021, 2018); Ilchmann et al. (2002). Here we do not allow for an arbitrary funnel boundary in order to be able to change its shape by means of a differential equation. However, we allow to prescribe the "asymptotic shape" $\psi(t) = ae^{-bt} + c$ under inactive saturation, that is the positive parameters a, b, c can be chosen as desired.

2. FUNNEL CONTROL STRUCTURE

We introduce the following input-constrained funnel controller for systems (1), (2).

$$\begin{aligned}
e_1(t) &= e(t) = y(t) - y_{\mathrm{ref}}(t),\\
e_{i+1}(t) &= e^{(i)}(t) + k_i(t)e_i(t), \quad i = 1, \dots, r-1,\\
k_i(t) &= \Big(1 - \tfrac{\|e_i(t)\|^2}{\psi_i(t)^2}\Big)^{-1}, \quad i = 1, \dots, r,\\
\dot{\psi}_i(t) &= p_i\psi_{i+1}(t) - \alpha_i\psi_i(t) + \beta_i - p_i\tfrac{\beta_{i+1}}{\alpha_{i+1}},\\
\psi_i(0) &= \psi_i^0, \qquad\qquad i = 1, \dots, r-1,\\
\dot{\psi}_r(t) &= -\alpha_r\psi_r(t) + \beta_r + \psi_r(t)\tfrac{\kappa(v(t))}{\|e_r(t)\|},\\
\psi_r(0) &= \psi_r^0,\\
\kappa(v(t)) &= \|v(t) - \mathrm{sat}(v(t))\|,\\
v(t) &= N\big(k_r(t)\big)e_r(t)
\end{aligned} \tag{3}$$

with the controller design parameters

$$\begin{aligned}
&\alpha_1 > \alpha_2 > \ldots > \alpha_r > 0,\ p_i > 1 \ \text{ for } i = 1, \dots, r-1,\\
&\beta_i > 0,\ \psi_i^0 > \frac{\beta_i}{\alpha_i} \ \text{ for } i = 1, \dots, r,\\
&N \in C(\mathbb{R}_{\geq 0}, \mathbb{R}) \ \text{ a surjection.}
\end{aligned} \tag{4}$$

Furthermore in (3) we assume that the instantaneous values of the output $y(t)$ and its derivatives $\dot{y}(t), \dots, y^{(r-1)}(t)$ are available for feedback, thus (3) is a dynamic output derivative feedback controller.

The first three equations of the controller (3) are basically a combination of the two designs from Berger et al. (2021, 2018), appended by the dynamics for the funnel boundaries in the subsequent three equations. This contrasts classical funnel control approaches, where the performance funnels are always prescribed a priori. Here, they are determined by a dynamical system, which is influenced by the input and an auxiliary error variable. Since the funnel functions are then used to determine these quantities in turn, a feedback structure arises, for which we seek to prove existence of global solutions.

The surjective function N in (4) serves the purpose of accommodating for possibly unknown control directions. With its help the controller is able to "probe" for the appropriate sign of the control signal. For more details see also (Berger et al., 2021, Rem. 1.8).

The distinguishing feature of the novel control design (3) is that it is feasible under arbitrary input constraints (2). The controller (3) always guarantees the evolution of the tracking error within a performance funnel, whose boundary is determined by a dynamic part of the controller as mentioned above. If the saturation is not active, then (asymptotically) the funnel boundary is of the form $\psi(t) = ae^{-bt} + c$ with positive design parameters a, b, c; if the saturation is active, i.e., $v(t) \neq \mathrm{sat}(v(t))$, then the boundary is widened by the dynamics of the controller in order to guarantee the input constraints. After a period of active saturation, the boundary recovers to its prescribed shape exponentially fast.

We emphasize that the controller (3) introduces several possible singularities in the closed-loop differential equation. In order to prove the existence of a global solution, it must be ensured that $\|e_i(t)\| \leq \varepsilon_i\psi_i(t)$ for some $\varepsilon_i \in (0,1)$ and that $\kappa(v(t)) = 0$ whenever $\|e_r(t)\| < \delta$ for some $\delta > 0$. Furthermore, compared to classical funnel control approaches as in Berger et al. (2021, 2018), the funnel boundaries ψ_i are not prescribed here, and in particular it is not known a priori that they are bounded. Hence, solutions may potentially get unbounded in finite time, i.e., exhibit a blow-up. Therefore, the feasibility proof of the control design is a highly nontrivial task.

3. FUNNEL CONTROL – MAIN RESULTS

In this section we show that the application of the funnel controller (3) to a system (1) under input constraints (2) leads to a closed-loop initial-value problem which has a global solution. By a solution of (1), (2), (3) on $[-h, \omega)$ we mean a tuple of functions $(y, \psi_1, \dots, \psi_r) \in C^{r-1}([-h,\omega), \mathbb{R}^m) \times C([-h,\omega), \mathbb{R})^r$ with $\omega \in (0, \infty]$, which satisfies $y|_{[-h,0]} = y^0$, $\psi_i(0) = \psi_i^0$ for all $i = 1, \dots, r$ and $(y^{(r-1)}, \psi_1, \dots, \psi_r)|_{[0,\omega)}$ is locally absolutely continuous and satisfies the differential equations in (1) and (3) with u defined by (2), (3) for almost all $t \in [0, \omega)$; $(y, \psi_1, \dots, \psi_r)$ is called maximal, if it has no right extension that is also a solution.

Next we present the main result.

Theorem 3. Consider a system (1) with $(d, f, T) \in \mathcal{N}^{m,r}$ for $m, r \in \mathbb{N}$, under input saturation (2) with saturation function sat that satisfies (P5). Let $y^0 \in C^{r-1}([-h, 0], \mathbb{R}^m)$ be the initial trajectory, $y_{\mathrm{ref}} \in C^r(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$ the reference signal and choose funnel control design parameters as in (4). Set $e = y - y_{\mathrm{ref}}$ and assume that the instantaneous values $e(t), \dot{e}(t), \dots, e^{(r-1)}(t)$ are available for feedback and satisfy, using the variables $e_1, \dots, e_r$ defined in (3), that

$$\forall\, i = 1, \dots, r: \ \|e_i(0)\| < \psi_i^0. \tag{5}$$

Then the funnel controller (3) applied to (1), (2) yields an initial-value problem which has a solution, every solution can be maximally extended and every maximal solution $(y, \psi_1, \dots, \psi_r) : [-h, \omega) \to \mathbb{R}^{m+r}$, $\omega \in (0, \infty]$, has the following properties:

(1) global existence: $\omega = \infty$;
(2) the functions $e_1, \dots, e_r$ evolve in their respective performance funnels in the sense:
$$\forall\, i = 1, \dots, r-1\ \exists\, \varepsilon_i \in (0,1)\ \forall\, t \geq 0: \\ \|e_i(t)\| \leq \varepsilon_i\psi_i(t) \quad \text{and} \quad \|e_r(t)\| < \psi_r(t);$$
(3) if the saturation is not active on some interval $[t_0, t_1) \subseteq \mathbb{R}_{\geq 0}$ with $t_1 \in (t_0, \infty]$, i.e., $v(t) = \mathrm{sat}(v(t))$ for all $t \in [t_0, t_1)$, then the performance funnels exponentially recover to their prescribed shape, i.e.,
$$\psi_i(t) \leq \tfrac{\beta_i}{\alpha_i} + \sum\nolimits_{j=i}^{r} \mu_j(t_0)\nu_{ij}e^{-\alpha_j(t-t_0)}$$

for all $i = 1, \dots, r$ and all $t \in [t_0, t_1)$, where $\mu_i(t_0) := \psi_i(t_0) - \frac{\beta_i}{\alpha_i}$, $\nu_{ii} := 1$ and $\nu_{ij} := \prod_{k=i}^{j-1} \frac{p_k}{\alpha_k - \alpha_j}$ for $i = 1, \dots, r$ and $j = i+1, \dots, r$.

We stress that although Theorem 3 provides the existence of a global solution of the closed-loop system, it cannot be concluded that the funnel boundaries $\psi_1, \dots, \psi_r$ are bounded in general. However, statement (iii) provides that *a posteriori* the funnel boundaries recover to their prescribed shape on any interval where the saturation is not active; in particular, if $t_1 = \infty$, then they are bounded.

Nevertheless, it is possible to show global boundedness of $\psi_1, \dots, \psi_r$ for sufficiently large saturation bound, i.e., $\mathrm{sat}(v) = v$ for all $v \in \mathbb{R}^m$ with $\|v\| \leq M$ and $M > 0$ sufficiently large. For this we require additional assumptions, i.e., a bounded reference signal with bounded derivatives and the system class $\mathcal{N}_{\mathrm{BIR}}^{m,r}$ considered for funnel control in Berger et al. (2021). We do not recall the precise definition of $\mathcal{N}_{\mathrm{BIR}}^{m,r}$ here, which can be found in (Berger et al., 2021, Def. 1.5), but only highlight the differences with $\mathcal{N}^{m,r}$: For $\mathcal{N}_{\mathrm{BIR}}^{m,r}$ property (P3) needs to hold with "$\tau = \infty$" (and hence becomes a bounded-input, bounded-output stability property) and system (1) needs to satisfy an additional high-gain property.

Theorem 4. Consider a system (1) with $(d, f, T) \in \mathcal{N}_{\mathrm{BIR}}^{m,r}$ for $m, r \in \mathbb{N}$. Choose funnel control design parameters as in (4), $\varepsilon \in (0,1)$ and $K > 0$. Then there exists $M > 0$ (depending on ε and K) such that

- for all saturation functions sat which satisfy (P5) with $\theta = M$,
- for all $y^0 \in C^{r-1}([-h, 0], \mathbb{R}^m)$ with $\|e_i(0)\| \leq \varepsilon \psi_i^0$, $i = 1, \dots, r$, and
- for all $y_{\mathrm{ref}} \in W^{r,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$ with $\|y_{\mathrm{ref}}^{(i)}\|_\infty \leq K$, $i = 0, \dots, r$,

there exists a solution $(y, \psi_1, \dots, \psi_r) : [-h, \omega) \to \mathbb{R}^{m+r}$, $\omega \in (0, \infty]$, of (1), (2), (3) which can be maximally extended to a global solution (i.e., $\omega = \infty$) that satisfies

(1) $y \in W^{r,\infty}([-h, \infty), \mathbb{R}^m)$ and $\psi_i \in L^\infty(\mathbb{R}_{\geq 0}, \mathbb{R})$ for $i = 1, \dots, r$;
(2) $k_i \in L^\infty(\mathbb{R}_{\geq 0}, \mathbb{R})$ for $i = 1, \dots, r$ and $v \in L^\infty(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$ with $\|v(t)\| \leq M$ for all $t \geq 0$, for the quantities defined in (3).

We stress that Theorem 4 provides an explicit relation between the initial values and the saturation bound M. The initial values $y(0), \dot{y}(0), \dots, y^{(r-1)}(0)$ are essentially confined to a bounded set, the size of which is quantified by $\varepsilon \in (0,1)$, for which the relations $\|e_i(0)\| \leq \varepsilon \psi_i^0$ hold for $i = 1, \dots, r$. If ε is made smaller, allowing only a smaller set of initial values, then it is also possible to choose a smaller saturation bound M in general (although $M \geq M^* > 0$ even for $\varepsilon \to 0$ and $K \to 0$, where M^* depends on the system and controller parameters).

REFERENCES

Bechlioulis, C.P. and Rovithakis, G.A. (2008). Robust adaptive control of feedback linearizable MIMO nonlinear systems with prescribed performance. *IEEE Trans. Autom. Control*, 53(9), 2090–2099.

Bechlioulis, C.P. and Rovithakis, G.A. (2014). A low-complexity global approximation-free control scheme with prescribed performance for unknown pure feedback systems. *Automatica*, 50(4), 1217–1226.

Berger, T., Ilchmann, A., and Ryan, E.P. (2021). Funnel control of nonlinear systems. *Math. Control Signals Syst.*, 33, 151–194.

Berger, T., Lê, H.H., and Reis, T. (2018). Funnel control for nonlinear systems with known strict relative degree. *Automatica*, 87, 345–357.

Berger, T. and Rauert, A.L. (2020). Funnel cruise control. *Automatica*, 119, Article 109061.

Berger, T. and Reis, T. (2014). Zero dynamics and funnel control for linear electrical circuits. *J. Franklin Inst.*, 351(11), 5099–5132.

Cheng, C., Zhang, Y., and Liu, S.Y. (2019). Neural observer-based adaptive prescribed performance control for uncertain nonlinear systems with input saturation. *Neurocomputing*, 370, 94–103.

Hackl, C.M. (2015). Current PI-funnel control with anti-windup for synchronous machines. In *Proc. 54th IEEE Conf. Decis. Control, Osaka, Japan*, 1997–2004.

Hackl, C.M. (2017). *Non-identifier Based Adaptive Control in Mechatronics–Theory and Application*, volume 466 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Cham, Switzerland.

Hackl, C.M., Hopfe, N., Ilchmann, A., Mueller, M., and Trenn, S. (2013). Funnel control for systems with relative degree two. *SIAM J. Control Optim.*, 51(2), 965–995.

Hopfe, N., Ilchmann, A., and Ryan, E.P. (2010a). Funnel control with saturation: linear MIMO systems. *IEEE Trans. Autom. Control*, 55(2), 532–538.

Hopfe, N., Ilchmann, A., and Ryan, E.P. (2010b). Funnel control with saturation: nonlinear SISO systems. *IEEE Trans. Autom. Control*, 55(9), 2177–2182.

Ilchmann, A., Ryan, E.P., and Sangwin, C.J. (2002). Tracking with prescribed transient behaviour. *ESAIM: Control, Optimisation and Calculus of Variations*, 7, 471–493.

Ilchmann, A. and Trenn, S. (2004). Input constrained funnel control with applications to chemical reactor models. *Syst. Control Lett.*, 53(5), 361–375.

Li, S. and Xiang, Z.R. (2018). Adaptive prescribed performance control for switched nonlinear systems with input saturation. *Int. J. Systems Sci.*, 49(1), 113–123.

Liberzon, D. and Trenn, S. (2013). The bang-bang funnel controller for uncertain nonlinear systems with arbitrary relative degree. *IEEE Trans. Autom. Control*, 58(12), 3126–3141.

Pomprapa, A., Weyer, S., Leonhardt, S., Walter, M., and Misgeld, B. (2015). Periodic funnel-based control for peak inspiratory pressure. In *Proc. 54th IEEE Conf. Decis. Control, Osaka, Japan*, 5617–5622.

Wang, Y., Hu, J., Li, J., and Liu, B. (2019). Improved prescribed performance control for nonaffine pure-feedback systems with input saturation. *Int. J. Robust & Nonlinear Control*, 29, 1769–1788.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Funnel control of linear systems under output measurement losses ⋆

Thomas Berger * **Lukas Lanza** *

* *Institut für Mathematik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Germany (e-mail: thomas.berger@math.upb.de, lanza@math.upb.de).*

Abstract: We consider tracking control of linear minimum phase systems with known arbitrary relative degree which are subject to possible output measurement losses. We provide a control law which guarantees the evolution of the tracking error within a (shifted) prescribed performance funnel whenever the output signal is available. The result requires a maximal duration of measurement losses and a minimal time of measurement availability, which both strongly depend on the internal dynamics of the system, and are derived explicitly.

Keywords: linear systems, funnel control, output tracking, measurement losses, minimum phase
AMS subject classifications: 93B52, 93C05, 93C40

1. INTRODUCTION

We study output tracking for linear minimum phase systems with arbitrary relative degree under possible output measurement losses. Such phenomena are of significant practical relevance whenever signals are transmitted over large distances or via digital communication networks and may hence be prone to signal losses or package dropouts. In the presence of output measurement losses the performance of closed-loop control strategies may seriously deteriorate and even lead to instability. In the present paper we present a reliable strategy for linear systems which is able to guarantee a prescribed margin for the tracking error and after any period of possible output measurement losses it is able to recapture the error within this time-varying margin by appropriately shifting it.

Output measurement losses are typically considered within the framework of networked control systems, see e.g. García-Rivera and Barreiro (2007); Wang and Yang (2009); Cloosterman et al. (2010); Nešić and Teel (2004). Within this approach, event-triggered controllers have been designed in order to guarantee global asymptotic stability, see Lehmann and Lunze (2012); Blind and Allgöwer (2014); Linsenmayer et al. (2019) for linear systems and Wang and Lemmon (2011); Dolk and Heemels (2017) for nonlinear systems. H_∞ control approaches have been considered in Gao and Chen (2008); Tang et al. (2016) and model predictive control in de la Pena and Christofides (2007); Lješnjanin et al. (2014). However, as far as the authors are aware, tracking control with prescribed performance bounds for the tracking error has not yet been considered. To achieve this, in the present paper we use the methodology of funnel control.

The concept of funnel control goes back to the seminal work Ilchmann et al. (2002), see also the survey in Berger et al. (2021c). The funnel controller proved to be the appropriate tool for tracking problems in various applications such as control of industrial servo-systems Hackl (2017) and underactuated multibody systems Berger et al. (2021b, 2019), control of electrical circuits Berger and Reis (2014); Senfelds and Paugurs (2014), control of peak inspiratory pressure Pomprapa et al. (2015), adaptive cruise control Berger and Rauert (2020) and even the control of infinite-dimensional systems such as a boundary controlled heat equation Reis and Selig (2015), a moving water tank Berger et al. (2022) and defibrillation processes of the human heart Berger et al. (2021a).

The novel funnel control design that we present in this paper relies on an intrinsic "availability function" which encodes (as a binary value) whether the output measurement is available at some time instant, or if the measurement is lost. As a consequence, no *a priori* information about the time instants where the measurement is lost or recaptured is necessary. Then the basic idea for the control design is simply to employ a classical funnel controller on each interval where the output is available, set the input to zero when it is not available and restart the controller when the output signal is received again. Because we restrict ourselves to linear systems no blow-up may occur when the input is zero. The crucial obstacle in the feasibility proof of the control design in our main result Theorem 5 is to show that the resulting control input in the closed-loop system is globally bounded. To this end, we require appropriate assumptions on the maximal duration of measurement losses and the minimal time of measurement availability, which we summarize in Section 1.2. The bounds for these durations essentially depend on the internal dynamics of the system – if the internal dynamics are absent, no restrictions must be made. However, if they are present a key step is to find an invariant set for the internal dynamics and to choose the initial width of the performance funnel large enough – this is elaborated in Section 1.3.

⋆ Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 362536361.

1.1 Nomenclature

We use the following notation, where $I \subseteq \mathbb{R}$ denotes an interval and $\mathbb{R}_{\geq 0} := [0, \infty)$. $\mathbb{N}$ is the set of positive integers; $\mathbb{C}_- := \{ z \in \mathbb{C} \mid \mathrm{Re} z < 0 \}$; $\|x\| := \sqrt{x^\top x}$ is the Euclidean norm of $x \in \mathbb{R}^n$; $\mathbf{Gl}_n(\mathbb{R})$ is the set of invertible matrices $A \in \mathbb{R}^{n\times n}$; for $A \in \mathbf{Gl}_n(\mathbb{R})$ we write $A > 0$ ($A < 0$) if A is positive (negative) definite; $\sigma(A) \subseteq \mathbb{C}$ is the spectrum of a matrix $A \in \mathbb{R}^{n\times n}$; $\mathcal{L}^\infty(I;\mathbb{R}^p)$ is the Lebesgue space of measurable and essentially bounded functions $f : I \to \mathbb{R}^p$ with norm $\|f\|_\infty := \operatorname{ess\,sup}_{t\in I} \|f(t)\|$; $\mathcal{W}^{k,\infty}(I;\mathbb{R}^p)$ is the Sobolev space of k-times weakly differentiable functions $f : I \to \mathbb{R}^p$ such that $f, \dots, f^{(k)} \in \mathcal{L}^\infty(I;\mathbb{R}^p)$; $\mathcal{C}^k(I;\mathbb{R}^p)$ is the set of k-times continuously differentiable functions $f : I \to \mathbb{R}^p$, $\mathcal{C}(I;\mathbb{R}^p) = \mathcal{C}^0(I;\mathbb{R}^p)$; $f|_J$ is the restriction of $f : I \to \mathbb{R}^n$ to $J \subseteq I$.

1.2 System class

We consider linear systems of the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad x(0) = x^0 \in \mathbb{R}^n, \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B, C^\top \in \mathbb{R}^{n\times m}$; in particular, the dimensions of the input $u(t)$ and the output $y(t)$ coincide. We assume that the system has strict relative degree $r \in \mathbb{N}$, i.e., $CA^kB = 0$ for all $k = 0, \dots, r-2$, and $\Gamma := CA^{r-1}B \in \mathbf{Gl}_m(\mathbb{R})$. Then, a straightforward generalization of (Ilchmann and Wirth, 2013, Thm. 3) yields that there exist $R_i \in \mathbb{R}^{m\times m}$, $S, P^\top \in \mathbb{R}^{m\times(n-rm)}$ and $Q \in \mathbb{R}^{(n-rm)\times(n-rm)}$ such that system (1) is equivalent to

$$\begin{aligned} y^{(r)}(t) &= \sum_{i=1}^{r} R_i y^{(i-1)}(t) + S\eta(t) + \Gamma u(t), \\ \dot{\eta}(t) &= Q\eta(t) + Py(t) \end{aligned} \tag{2}$$

with initial conditions

$$\begin{aligned} (y(0), \dots, y^{(r-1)}(0)) &= (y^0_0, \dots, y^0_{r-1}) \in \mathbb{R}^{rm}, \\ \eta(0) &= \eta^0 \in \mathbb{R}^{n-rm}. \end{aligned}$$

We introduce the system class under consideration.

Definition 1. For $r, m \in \mathbb{N}$ a system (2) belongs to the system class $\Sigma_{r,m}$, if

(i) the high-gain matrix $\Gamma \in \mathbf{Gl}_m(\mathbb{R})$ is sign definite [1]; w.l.o.g. we assume $\Gamma + \Gamma^\top > 0$,
(ii) the system is minimum phase, i.e., $\sigma(Q) \subseteq \mathbb{C}_-$.

We write $(A, B, C) \in \Sigma_{r,m}$.

We record the following, the proof of which is straightforward.

Lemma 2. For $L \in \mathbb{R}^{p\times p}$ with $\sigma(L) \subseteq \mathbb{C}_-$ there exists $0 < K = K^\top$ such that $KL + L^\top K = -I_p$, and

$$\forall\, t \geq 0 : \ \|e^{Lt}\| \leq \sqrt{\|K^{-1}\|\|K\|}\, e^{-\frac{1}{2\|K\|}t}.$$

In virtue of Lemma 2, for Q from (2) let

$$M := \sqrt{\|K^{-1}\|\|K\|}, \ \mu := \frac{1}{2\|K\|}, \tag{3}$$

where $KQ + Q^\top K = -I_{n-rm}$. If $n - rm = 0$, then we set $M := 0$ and $\mu := 1$.

[1] That is, for any $v \in \mathbb{R}^m$ we have $v^\top \Gamma v = 0$ if, and only if, $v = 0$.

Since we consider situations where the output measurement signal may be lost for some time, we propose assumptions relating the maximal duration of measurement losses and minimal time of measurement availability. The package dropouts in the system and the accompanying lost information of the measurements $y(t)$ are not assumed to happen in *a priori* known time intervals. We only assume that it is possible to determine, at every time instant t, whether the measurement of $y(t)$ is available or not; if the availability is not certain, then it should be rendered "unavailable" (this also encompasses the situation that, after a dropout, the availability of the measurement is only determined with some delay). Based on this we define an "availability function"

$$a(t) = \begin{cases} 1, & \text{measurement of } y(t) \text{ available,} \\ 0, & \text{measurement of } y(t) \text{ not available.} \end{cases} \tag{4}$$

Let (t_k^-), (t_k^+) be sequences with $t_k^\pm \nearrow \infty$ and $t_k^- < t_k^+ < t_{k+1}^- < t_{k+1}^+$ such that

$$\begin{aligned} \{\, t \geq 0 \mid a(t) = 1 \,\} &= \bigcup_{k\in\mathbb{N}} (t_k^+, t_{k+1}^-], \\ \{\, t \geq 0 \mid a(t) = 0 \,\} &= \bigcup_{k\in\mathbb{N}} (t_k^-, t_k^+], \end{aligned} \tag{5}$$

that is, on the interval $(t_k^+, t_{k+1}^-]$ the signal is available, and on the interval $(t_k^-, t_k^+]$ the signal is not available. Note, that it is also possible that both sequences contain only finitely many points, then either $a(t) = 1$ for $t \geq t_N^+$ or $a(t) = 0$ for $t \geq t_N^-$ for some $N \in \mathbb{N}$. Now, we assume the following on the maximal duration of measurement losses and the minimal time of measurement availability.

Assumption 1. Let $p := \|P\|$, $s := \|S\|$ and $\beta := 1 + \frac{spM}{\mu} + \sum_{i=1}^r \|R_i\|$ be given by the system parameters, M, μ from (3) and q, A_r be the constants introduced in Section 1.3. The signal is lost for at most $\Delta > 0$, i.e., for $t_k^\pm$ as in (5) we have $|t_k^- - t_k^+| \leq \Delta$ for all $k \in \mathbb{N}$, such that for some $\kappa \geq 2$ and $\theta > s$ we have that Δ satisfies

$$spM\Delta^2 e^{\beta\Delta} \leq 1, \tag{Δ_1}$$

$$pM^2\Delta e^{\beta\Delta} \leq \frac{q}{A_r} \cdot \frac{\mu(\kappa-1)}{2\kappa\theta}. \tag{Δ_2}$$

Assumption 2. The signal is available for at least $\delta > 0$, i.e., for $t_k^\pm$ as in (5) we have $|t_k^+ - t_{k+1}^-| \geq \delta$ for all $k \in \mathbb{N}$, such that for Δ, β, κ, θ from Assumption 1 and M, μ from (3) we have that δ satisfies

$$e^{\mu\delta} \geq 2\kappa M\big(M + p\Delta e^{\beta\Delta}(1 + sM^2\Delta)\big), \tag{δ_1}$$

$$e^{\mu\delta} \geq 2\frac{\kappa}{\theta}(1 + sM^2). \tag{δ_2}$$

Remark 3. For systems with trivial internal dynamics (the second equation in (2) is not present) Assumptions 1 & 2 are much weaker. In this case we have $p = 0$, $s = 0$ and $M = 0$ with which the inequalities (Δ_1), (Δ_2) and (δ_1), (δ_2) are always satisfied (for $\theta = 2\kappa$). Hence, arbitrary $\Delta > 0$ and $\delta > 0$ are possible so that $|t_k^- - t_k^+| \leq \Delta$ and $|t_k^+ - t_{k+1}^-| \geq \delta$ for all $k \in \mathbb{N}$. So the only (implicit) requirement is that the sequence $(|t_k^- - t_k^+|)$ is bounded.

1.3 Control objective, design parameters and feedback law

Control objective We aim to find a control scheme which achieves tracking of a given reference trajectory

with prescribed transient behavior of the error, where the measurement output is subject to dropouts. To be more precise, for a system (2) with $(A, B, C) \in \Sigma_{r,m}$ and a given reference signal $y_{\mathrm{ref}} \in \mathcal{W}^{r,\infty}(\mathbb{R}_{\geq 0}; \mathbb{R}^m)$ the output y tracks the reference in the sense that the error $e(\cdot) := y(\cdot) - y_{\mathrm{ref}}(\cdot)$, whenever the measurement of y is available to the controller, evolves within a prescribed *performance funnel*

$$\mathcal{F}_\varphi := \{\, (t, e) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^m \mid \varphi(t)\|e\| < 1 \,\},$$

where φ belongs to the set Φ of monotonically increasing functions introduced below.

$$\Phi := \left\{ \phi \in \mathcal{C}^1(\mathbb{R}_{\geq 0}; \mathbb{R}) \;\middle|\; \begin{array}{l} \forall\, t_2 \geq t_1 \geq 0 : \\ 0 < \phi(t_1) \leq \phi(t_2), \\ \exists\, d > 0\ \forall\, t \geq 0 : \\ |\dot{\phi}(t)| \leq d(1 + \phi(t)) \end{array} \right\}.$$

The performance funnel $\mathcal{F}_\varphi$ joins the two objectives of $e(t)$ approaching zero with prescribed transient behavior and asymptotic accuracy. Its boundary is given by the reciprocal of φ. We stress that φ may be unbounded and in this case (and if no measurement losses occur for $t \geq T$ for some $T > 0$) asymptotic tracking may be achieved, i.e., $\lim_{t\to\infty} e(t) = 0$.

Design parameters In order to formulate the control law, which achieves the control objective, we introduce the following design parameters. *Step 1.* Choose $q \in (0, 1)$ and define the bijection $\alpha : [0, 1) \to [1, \infty)$ via $\alpha(s) = 1/(1-s)$. For $k \geq 0$ define the function

$$A_k(s) = \sum_{j=0}^{k} s^j,$$

and set

$$A_r := A_r(\alpha(q^2)).$$

Step 2. For $\Delta, \delta, p, s, \beta, \kappa, \theta$ from Assumptions 1 & 2, respectively, $x_{\mathrm{ref}}(\cdot) := (y_{\mathrm{ref}}(\cdot), \dot{y}_{\mathrm{ref}}(\cdot), \ldots, y_{\mathrm{ref}}^{(r-1)}(\cdot))$, and M, μ from (3) choose $\eta^* > 0$ with

$$\eta^* \geq \max \left\{ \begin{array}{l} \frac{p}{\mu}\|y_{\mathrm{ref}}\|_\infty e^{\mu\delta}, \|x_{\mathrm{ref}}\|_\infty e^{\mu\delta}, \\ \dfrac{\|x_{\mathrm{ref}}\|_\infty (1 + e^{\beta\Delta}) e^{\mu\delta - \beta\Delta}}{\Delta} \end{array} \right\}, \tag{6}$$

and set

$$E := \theta\Delta e^{\beta\Delta}\eta^* > 0.$$

Step 3. Let $\varphi_0 \in \Phi$ such that

$$\varphi_{0,\min} := \tfrac{2\kappa p M^2}{\mu(\kappa-1)\eta^*} \leq \varphi_0(0) \leq \tfrac{q}{A_r E} =: \varphi_{0,\max}, \tag{ϕ_1}$$

which is possible by (Δ_2).

Step 4. Now, we choose some additional constants which are necessary to exploit (Berger et al., 2021c, Cor. 1.10). Let $\hat{\alpha}^\dagger(z) = z/(1+z)$, and define $\tilde{\alpha}(s) := 2s\alpha'(s) + \alpha(s) = (1+s)/(1-s)^2$. Further, let $\mu_0 := \operatorname{ess\,sup}_{t\geq 0}(|\dot{\varphi}_0(t)|/\varphi_0(t))$ which by properties of Φ satisfies $\mu_0 \leq \frac{d(1+\varphi_0(0))}{\varphi_0(0)}$ for some $d > 0$. Then, in virtue of (Berger et al., 2021c, Eq. (12)), for $k = 1, \ldots, r-1$ we recursively define the constants $c_0 = 0$ and

$$\begin{aligned} e_1^0 &:= \varphi_0(0)e(0), \\ c_1 &:= \max\{\|e_1^0\|^2, \hat{\alpha}^\dagger(1+\mu_0), q^2\}^{1/2} < 1, \\ \mu_k &:= 1 + \mu_0\big(1 + c_{k-1}\alpha(c_{k-1}^2)\big) \\ &\quad + \tilde{\alpha}(c_{k-1}^2)\big(\mu_{k-1} + c_{k-1}\alpha(c_{k-1}^2)\big), \\ e_k^0 &:= \varphi_0(0)e^{(k-1)}(0) + \alpha(\|e_{k-1}^0\|^2)e_{k-1}^0, \\ c_k &:= \max\{\|e_k^0\|^2, \hat{\alpha}^\dagger(\mu_k), q^2\}^{1/2} < 1, \end{aligned}$$

where $e^{(i)}(0) = y_i^0 - y_{\mathrm{ref}}^{(i)}(0)$ for $i = 0, \ldots, r-1$, and set

$$C := \textstyle\sum_{i=1}^{r-1} c_i + c_{i-1}\alpha(c_{i-1}^2) + (1 + c_{r-1}\alpha(c_{r-1}^2)).$$

Step 5. We refine the function $\varphi_0 \in \Phi$ satisfying (ϕ_1) such that for an intermediate $\rho \in (0, \delta)$

$$\varphi_0(\rho) \geq \max\left\{ \tfrac{Ce^{\mu\delta}}{\eta^*}, \tfrac{C\,e^{\mu\delta}}{\Delta\eta^*} \right\}. \tag{ϕ_2}$$

Remark 4. We note that the purpose of the constant q chosen in *Step 1* of the design procedure is to determine the initial width of the performance funnel, described by the upper bound for $\varphi_0(0)$ in (ϕ_1). Then again, condition (ϕ_2) ensures that its width (and hence the tracking error) is not too large before the signal possibly vanishes the next time.

Feedback law The idea for the controller design is to choose a funnel function $\varphi_0 \in \Phi$ (as in the previous subsection) which is reset whenever $a(t) = 0$. Then, as soon as $a(t^*) = 1$ for some $t^* \geq 0$ and the measurement is available again, the funnel controller from Berger et al. (2021c) is restarted with $\varphi(t) = \varphi_0(t - t^*)$ so that $\varphi(t^*) > 0$ and the performance funnel is sufficiently large at t^* to ensure applicability of (Berger et al., 2021c, Thm. 1.9). For feasibility we assume that the availability function $a(\cdot)$ from (4) is left-continuous and has only finitely many jumps in each compact interval. With this, and recalling $\alpha(s) = 1/(1-s)$, we introduce the following control law for systems (2) under possible output measurement losses:

$$\begin{aligned} \tau(t) &= \begin{cases} t, & a(t) = 0, \\ \tau(t-), & a(t) = 1, \end{cases} \\ \varphi(t) &= \begin{cases} 0, & a(t) = 0, \\ \varphi_0(t - \tau(t)), & a(t) = 1, \end{cases} \\ e_1(t) &= \varphi(t)e(t) = \varphi(t)\big(y(t) - y_{\mathrm{ref}}(t)\big), \\ e_{i+1}(t) &= \varphi(t)e^{(i)}(t) + \alpha(\|e_i(t)\|^2)e_i(t), i = 1, \ldots, r-1, \\ u(t) &= -a(t)\alpha(\|e_r(t)\|^2)e_r(t). \end{aligned} \tag{7}$$

Note that if $\Gamma + \Gamma^\top < 0$ the proposed control would read $u(t) = a(t)\alpha(\|e_r(t)\|^2)e_r(t)$.

If the output measurement is always available, i.e., $a(t) = 1$ for all $t \geq 0$, then the controller (7) coincides with that proposed in Berger et al. (2021c) and the existence of a global solution of the closed-loop system follows from the results presented there. Since it is not known *a priori* when output measurement losses occur, the funnel function φ cannot be globally defined in advance. Therefore, φ is defined online as part of the control law (7); it is equal to a shifted version of the reference funnel function φ_0 whenever measurements are available, and zero otherwise. Note that the loss of the system's output signal possibly introduces a discontinuity in the control signal.

2. MAIN RESULT

We show that the application of the funnel controller (7) to a system (2) under possible output measurement losses leads to a closed-loop initial-value problem which has a global solution. By a solution of (2), (7) on $[0, \omega)$ we mean a function $(y, \eta) \in \mathcal{C}^{r-1}([0,\omega), \mathbb{R}^m) \times \mathcal{C}([0,\omega), \mathbb{R}^{n-rm})$ with $\omega \in (0, \infty]$, which satisfies $(y(0), \dots, y^{(r-1)}(0)) = (y_0^0, \dots, y_{r-1}^0)$, $\eta(0) = \eta^0$ and $(y^{(r-1)}, \eta)|_{[0,\omega)}$ is locally absolutely continuous and satisfies (2) with u defined by (7) for almost all $t \in [0, \omega)$; (y, η) is called maximal, if it has no right extension that is also a solution.

Theorem 5. Consider a system (2) with $(A, B, C) \in \Sigma_{r,m}$ and initial values $(y_0^0, \dots, y_{r-1}^0) \in \mathbb{R}^{rm}$ and $\eta^0 \in \mathbb{R}^{n-rm}$. Let $y_{\text{ref}} \in \mathcal{W}^{r,\infty}(\mathbb{R}_{\geq 0}; \mathbb{R}^m)$, $a(\cdot)$ be as in (4) and choose design parameters η^* as in (6), and $\varphi_0 \in \Phi$ satisfying (ϕ_1),(ϕ_2). If the initial conditions

$$\forall\, i = 1, \dots, r \,:\, \|e_i(0)\| < 1, \quad \|\eta^0\| \leq \eta^*$$

are satisfied, then the control scheme (7) applied to system (2) yields an initial value problem which has a solution, every solution can be extended to a maximal solution and every maximal solution $(y, \eta) : [0, \omega) \to \mathbb{R}^m \times \mathbb{R}^{n-rm}$ has the following properties:

(i) the solution is global, i.e., $\omega = \infty$,
(ii) the tracking error $e(t) = y(t) - y_{\text{ref}}(t)$ evolves within the funnel boundaries, i.e., for all $t \geq 0$ we have $\varphi(t)\|e(t)\| < 1$,
(iii) the input control signal is globally bounded, i.e., $u \in \mathcal{L}^\infty(\mathbb{R}_{\geq 0}; \mathbb{R}^m)$, and $y \in \mathcal{W}^{r,\infty}(\mathbb{R}_{\geq 0}; \mathbb{R}^m)$.

REFERENCES

Berger, T., Breiten, T., Puche, M., and Reis, T. (2021a). Funnel control for the monodomain equations with the FitzHugh-Nagumo model. *Journal of Differential Equations*, 286, 164–214.

Berger, T., Drücker, S., Lanza, L., Reis, T., and Seifried, R. (2021b). Tracking control for underactuated non-minimum phase multibody systems. *Nonlinear Dynamics*, 104(4), 3671–3699.

Berger, T., Ilchmann, A., and Ryan, E.P. (2021c). Funnel control of nonlinear systems. *Mathematics of Control, Signals, and Systems*, 33(1), 151–194.

Berger, T., Otto, S., Reis, T., and Seifried, R. (2019). Combined open-loop and funnel control for underactuated multibody systems. *Nonlinear Dynamics*, 95, 1977–1998.

Berger, T., Puche, M., and Schwenninger, F.L. (2022). Funnel control for a moving water tank. *Automatica*, 135, 109999.

Berger, T. and Rauert, A.L. (2020). Funnel cruise control. *Automatica*, 119, 109061.

Berger, T. and Reis, T. (2014). Zero dynamics and funnel control for linear electrical circuits. *J. Franklin Inst.*, 351(11), 5099–5132.

Blind, R. and Allgöwer, F. (2014). On the stabilizability of continuous-time systems over a packet based communication system with loss and delay. *IFAC Proceedings Volumes*, 47(3), 6466–6471.

Cloosterman, M., Hetel, L., van de Wouw, N., Heemels, W., Daafouz, J., and Nijmeijer, H. (2010). Controller synthesis for networked control systems. *Automatica*, 46(10), 1584–1594.

de la Pena, D.M. and Christofides, P.D. (2007). Lyapunov-based model predictive control of nonlinear systems subject to data losses. In *2007 American Control Conference*. IEEE.

Dolk, V. and Heemels, M. (2017). Event-triggered control systems under packet losses. *Automatica*, 80, 143–155.

Gao, H. and Chen, T. (2008). Network-based $\mathcal{H}_\infty$ output tracking control. *IEEE Transactions on Automatic Control*, 53(3), 655–667.

García-Rivera, M. and Barreiro, A. (2007). Analysis of networked control systems with drops and variable delays. *Automatica*, 43(12), 2054–2059.

Hackl, C.M. (2017). *Non-identifier Based Adaptive Control in Mechatronics–Theory and Application*, volume 466 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Cham, Switzerland.

Ilchmann, A. and Ryan, E.P. (2009). Performance funnels and tracking control. *Int. J. Control*, 82(10), 1828–1840.

Ilchmann, A., Ryan, E.P., and Sangwin, C.J. (2002). Tracking with prescribed transient behaviour. *ESAIM: Control, Optimisation and Calculus of Variations*, 7, 471–493.

Ilchmann, A. and Wirth, F. (2013). On minimum phase. *Automatisierungstechnik*, 12, 805–817.

Lehmann, D. and Lunze, J. (2012). Event-based control with communication delays and packet losses. *International Journal of Control*, 85(5), 563–577.

Linsenmayer, S., Dimarogonas, D.V., and Allgöwer, F. (2019). Periodic event-triggered control for networked control systems based on non-monotonic lyapunov functions. *Automatica*, 106, 35–46.

Lješnjanin, M., Quevedo, D.E., and Nešić, D. (2014). Packetized MPC with dynamic scheduling constraints and bounded packet dropouts. *Automatica*, 50(3), 784–797.

Nešić, D. and Teel, A. (2004). Input–output stability properties of networked control systems. *IEEE Transactions on Automatic Control*, 49(10), 1650–1667.

Pomprapa, A., Weyer, S., Leonhardt, S., Walter, M., and Misgeld, B. (2015). Periodic funnel-based control for peak inspiratory pressure. In *Proc. 54th IEEE Conf. Decis. Control, Osaka, Japan*, 5617–5622.

Reis, T. and Selig, T. (2015). Funnel control for the boundary controlled heat equation. *SIAM J. Control Optim.*, 53(1), 547–574.

Seifried, R. and Blajer, W. (2013). Analysis of servo-constraint problems for underactuated multibody systems. *Mech. Sci.*, 4, 113–129.

Senfelds, A. and Paugurs, A. (2014). Electrical drive DC link power flow control with adaptive approach. In *Proc. 55th Int. Sci. Conf. Power Electr. Engg. Riga Techn. Univ., Riga, Latvia*, 30–33.

Tang, Y., Gao, H., and Kurths, J. (2016). Robust $\mathcal{H}_\infty$ self-triggered control of networked systems under packet dropouts. *IEEE Transactions on Cybernetics*, 46(12), 3294–3305.

Wang, X. and Lemmon, M.D. (2011). Event-triggering in distributed networked control systems. *IEEE Transactions on Automatic Control*, 56(3), 586–601.

Wang, Y.L. and Yang, G.H. (2009). Time delay and packet dropout compensation for networked control systems: a linear estimation method. *International Journal of Control*, 83(1), 115–124.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Least squares moment matching for linear and nonlinear systems

Alberto Padoan *

* *Automatic Control Laboratory, ETH Zürich, Zürich, Switzerland (e-mail:apadoan@ethz.ch).*

Abstract: The model reduction problem by least squares moment matching is studied. A recent time-domain characterization of least squares moment matching for linear systems is used to define a notion of least squares moment matching for nonlinear systems. Models achieving least squares moment matching are shown to minimize an *a priori* error bound on the worst case r.m.s. gain of an error system with respect to a given family of signals, thus providing new insights on the linear theory.

Keywords: Model reduction, nonlinear systems, least squares, moment matching.

1. INTRODUCTION

Model reduction is art of approximating of a system while retaining its most essential properties (Antoulas, 2005). A classical solution to this problem relies on the notion of moment matching (Antoulas, 2005). For linear systems, the main idea is to use rational interpolation theory to ensure that the coefficients of the Laurent series expansion of the transfer functions of the original system and of its approximant coincide at given points of the complex plane up to a given order. Following the seminal contributions (Grimme, 1997; Gallivan et al., 2004, 2006), moment matching extended to nonlinear systems using invariance equations and steady-state responses (Astolfi, 2010). This, in turn, has led to the development of new model reduction methods for several classes of systems, including time-delay systems (Scarciotti and Astolfi, 2016) and systems with singularities (Padoan and Astolfi, 2019).

A significant limitation of methods based on moment matching is that the interpolation conditions are required to hold exactly, which, for some purposes, is an unnecessarily stringent assumption. Furthermore, methods based on moment matching do not provide *a priori* guarantees on the quality of approximation in general. Least squares moment matching provides a particularly interesting solution to both issues (Aguirre, 1992; Smith and Lucas, 1995; Gugercin and Antoulas, 2006), requiring that the interpolation conditions imposed by moment matching are satisfied only in a least squares sense. Least squares moment matching thus overcomes the issues mentioned above by minimizing an optimization criterion, which directly yields *a priori* error bounds and, under certain assumptions, guaranteed stability properties (Gugercin and Antoulas, 2006). For linear systems, there is a vast literature on least squares moment matching (Aguirre, 1992; Smith and Lucas, 1995; Gugercin and Antoulas, 2006; Mayo and Antoulas, 2007; Nakatsukasa et al., 2018), with deep connections to Padé approximation (Aguirre, 1992) and Prony's method for filter design (Gugercin and Antoulas, 2006). Yet, the notion of least squares moment matching does not have a nonlinear counterpart to date.

Goal of this work is to present a unifying notion of least squares moment matching for linear and nonlinear systems. The main ingredient of our approach is the formalism introduced in (Astolfi, 2010), where moments of nonlinear systems have been defined and characterized using tools from output regulation theory (Isidori, 1995). The starting point of our discussion is a new time-domain characterization of least squares moment matching for linear systems first presented in (Padoan, 2021), which relies on the solution of a constrained optimization problem involving a Sylvester equation. In close analogy with the results obtained in (Padoan, 2021) for linear systems, models achieving least squares moment matching are defined in terms of a constrained optimization problem involving an invariance equation and shown to minimize an *a priori* error bound on the worst case r.m.s. gain of an error system with respect to a given family of signals.

The remainder of this work is organized as follows. Section 2 is dedicated to the model reduction problem by least squares moment matching for linear systems. Section 3 provides a unifying notion of least squares moment matching and a direct nonlinear counterpart of the linear theory. Section 4 concludes this work with a summary and an outlook to future research directions.

Notation: $\mathbb{Z}_+$, $\mathbb{R}$ and $\mathbb{C}$ denote the set of non-negative integers, of real numbers, and of complex numbers, respectively. $\mathbb{C}_-$ denotes the set of complex numbers with negative real part. e_k denotes the vector with the k-th entry equal to one and all other entries equal to zero. I denotes the identity matrix. J_0 denotes the matrix with ones on the superdiagonal and zeros elsewhere. $J_{s^\star}$ denotes the Jordan block associated with the eigenvalue $s^\star \in \mathbb{C}$, *i.e.* $J_{s^\star} = s^\star I + J_0$. $\sigma(A)$ denotes the spectrum of the matrix $A \in \mathbb{R}^{n\times n}$. M^T denotes the transpose of the matrix $M \in \mathbb{R}^{p\times m}$. $\|\cdot\|_2$ and $\|\cdot\|_{2*}$ denote the Euclidean 2-norm on $\mathbb{R}^n$ and the corresponding dual norm (Boyd and Vandenberghe, 2004, p.637), respectively. Finally, $f^{(k)}(\cdot)$ denotes the derivative of order $k \in \mathbb{Z}_+$ of the function $f(\cdot)$, provided it exists, with $f^{(0)}(\cdot) = f(\cdot)$ by convention.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2. LINEAR SYSTEMS

Consider a system described by the equations

$$\dot{x} = Ax + Bu, \quad y = Cx, \tag{1}$$

in which $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$ and $C \in \mathbb{R}^{1\times n}$ are constant matrices, with transfer function defined as

$$W(s) = C(sI - A)^{-1}B.$$

The system (1) is assumed to be minimal, *i.e.* controllable and observable.

Definition 1. (Antoulas, 2005, p.345) The *moment of order* $k \in \mathbb{Z}_+$ of system (1) at $s^\star \in \mathbb{C}$, with $s^\star \notin \sigma(A)$, is defined as the complex number

$$\eta_k(s^\star) = \frac{(-1)^k}{k!} W^{(k)}(s^\star).$$

Given distinct *interpolation points* $\{s_i\}_{i=1}^N$, with $s_i \in \mathbb{C}$ and $s_i \notin \sigma(A)$, and the corresponding *orders of interpolation* $\{k_i\}_{i=1}^N$, with $k_i \in \mathbb{Z}_+$, model reduction by moment matching consists in finding a system

$$\dot{\xi} = F\xi + Gv, \quad \psi = H\xi, \tag{2}$$

where $\xi(t) \in \mathbb{R}^r$, $v(t) \in \mathbb{R}$, $\psi(t) \in \mathbb{R}$ and $F \in \mathbb{R}^{r\times r}$, $G \in \mathbb{R}^{r\times 1}$ and $H \in \mathbb{R}^{1\times r}$ are constant matrices, the transfer function of which

$$\hat{W}(s) = H(sI - F)^{-1}G$$

satisfies the *interpolation conditions*

$$\eta_j(s_i) = \hat{\eta}_j(s_i), \quad j \in \{0, \dots, k_i\},\ i \in \{1, \dots, N\}, \tag{3}$$

where $\eta_j(s_i)$ and $\hat{\eta}_j(s_i)$ denote the moments of order j of the systems (1) and (2) at s_i, respectively. The system (2) is referred to as a *model (of system* (1)*)* and is said to *achieve moment matching (at* $\{s_i\}_{i=1}^N$*)* if the interpolation conditions (3) hold (Antoulas, 2005, Chapter 11). Furthermore, if $r < n$, then (2) is said to be a *reduced order model (of system* (1)*)*.

We are interested in the following problem. Suppose that the number of interpolation conditions

$$\nu = \sum_{i=1}^{N} (k_i + 1)$$

is larger than the number of moments that can be matched, i.e. $\nu > 2r$ (Antoulas, 2005, Chapter 11). In this case, the interpolation conditions (3) give rise to an overdetermined system of equations which can be solved in a least squares sense, leading directly to the *model reduction problem by least squares moment matching*.

Problem 1. Consider system (1). Let $\{s_i\}_{i=1}^N$ be a set of distinct interpolation points, with $s_i \in \mathbb{C}$ and $s_i \notin \sigma(A)$, and let $\{k_i\}_{i=1}^N$ be the corresponding orders of interpolation, with $k_i \in \mathbb{Z}_+$. Let $\nu = \sum_{i=1}^N (k_i + 1)$ and $r \in \mathbb{Z}_+$, with $2r < \nu$. Find, if possible, a model (2) of order r which minimizes the index

$$\mathcal{J} = \sum_{i=1}^{N} \sum_{j=0}^{k_i} \left| \eta_j(s_i) - \hat{\eta}_j(s_i) \right|^2. \tag{4}$$

The model (2) is said to *achieve least squares moment matching (at* $\{s_i\}_{i=1}^N$*)* if it minimizes the index (4).

2.1 Least squares moment matching for linear systems

Following (Padoan, 2021), we begin our analysis with a characterization of least squares moment matching in terms of the solutions of the optimization problem

$$\begin{aligned} \text{minimize} \quad & \|(C\Pi - HP)T\|_{2*}^2 \\ \text{subject to} \quad & A\Pi + BL = \Pi S, \\ & FP + GL = PS, \\ & \sigma(F) \cap \sigma(S) = \emptyset, \end{aligned} \tag{5}$$

for a given non-singular matrix $T \in \mathbb{R}^{\nu\times\nu}$, where $F \in \mathbb{R}^{r\times r}$, $G \in \mathbb{R}^{r\times 1}$, $H \in \mathbb{R}^{1\times r}$ and $P \in \mathbb{R}^{r\times\nu}$ are the optimization variables, and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, $C \in \mathbb{R}^{1\times n}$, $S \in \mathbb{R}^{\nu\times\nu}$ and $L \in \mathbb{R}^{1\times\nu}$ are problem data. To this end, we first introduce some basic assumptions.

(A1) The matrix $S \in \mathbb{R}^{\nu\times\nu}$ is non-derogatory [1] and has characteristic polynomial

$$\chi_S(s) = \prod_{i=1}^{N} (s - s_i)^{k_i+1}. \tag{6}$$

(A2) The matrix $L \in \mathbb{R}^{1\times\nu}$ is such that the pair (S, L) is observable.

(A3) The matrix $T \in \mathbb{R}^{\nu\times\nu}$ is non-singular and such that

$$ST = TJ, \quad LT = \Lambda, \tag{7}$$

with $J = \operatorname{diag}(J_{s_1}, \dots, J_{s_N})$ and $\Lambda = [\, e_1^\mathsf{T} \ \cdots \ e_1^\mathsf{T} \,]$.

Theorem 2. Consider system (1) and the model (2). Suppose Assumptions (A1)-(A3) hold. Assume $\sigma(A) \cap \sigma(S) = \emptyset$. Then the model (2) achieves least squares moment matching at $\sigma(S)$ if and only if there exists a matrix $P \in \mathbb{R}^{r\times\nu}$ such that (F, G, H, P) is a solution of the optimization problem (5).

Least squares moment matching admits a nice interpretation in terms of the steady-state behavior of the error system

$$\dot{x} = Ax + Bu, \quad \dot{\xi} = F\xi + Gu, \quad e = Cx - H\xi, \tag{8}$$

in which $x(t) \in \mathbb{R}^n$, $\xi(t) \in \mathbb{R}^r$, $u(t) \in \mathbb{R}$, and $e(t) \in \mathbb{R}$. Specifically, suppose that both the system (1) and the model (2) are driven by a signal generator described by the equations

$$\dot{\omega} = S\omega, \quad \theta = L\omega, \tag{9}$$

with $\omega(t) \in \mathbb{R}^\nu$ and $\theta(t) \in \mathbb{R}$. Furthermore, suppose that all solutions of the signal generator (9) are periodic and that the steady-state output response [2] e_{ss} of the interconnected system (9)-(8), with $u = \theta$, is well-defined. Then achieving least squares moment matching corresponds to minimizing an upper bound of the *worst case r.m.s. gain* of the error system (8) with respect to the family of signals produced by the signal generator (9), defined as (Boyd and Barratt, 1991, p.98)

$$\gamma_{rms} = \sup_{\omega(\cdot)\in\mathcal{W}} \frac{\|e_{ss}\|_{rms}}{\|\omega\|_{rms}} \tag{10}$$

where $\|\omega\|_{rms}$ is the *r.m.s. value* of the signal $\omega(t) \in \mathbb{R}^\nu$, defined as (Boyd and Barratt, 1991, p.86)

$$\|\omega\|_{rms} = \left(\lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \|\omega(t)\|_2^2 \, dt \right)^{1/2}, \tag{11}$$

[1] A matrix is non-derogatory if its characteristic polynomials and its minimal polynomial coincide (Horn and Johnson, 1994, p.178).
[2] See (Isidori, 1995, Chapter 8).

provided the limit exists, while $\mathcal{W}$ is the family of signals produced by (9) with non-zero r.m.s. value.

Theorem 3. Consider system (1), the model (2) and the signal generator (9). Suppose Assumptions (A1)-(A3) hold. Assume that $\sigma(A) \cup \sigma(F) \subset \mathbb{C}_-$, $S + S^\mathsf{T} = 0$ and $(S, \omega(0))$ is controllable. Then the following statements hold.

(i) The steady-state output response of the interconnected system (9)-(8), with $u = \theta$, is well-defined and uniquely determined by the moments of the error system (8) at $\sigma(S)$.

(ii) The worst case r.m.s. gain of the error system (8) with respect to the family of signals produced by the signal generator (9) is well-defined and such that

$$\gamma_{rms} \leq \|C\Pi - HP\|_{2*}, \tag{12}$$

with $\Pi \in \mathbb{R}^{n\times\nu}$ and $P \in \mathbb{R}^{r\times\nu}$ the (unique) solutions of the Sylvester equations

$$A\Pi + BL = \Pi S. \tag{13}$$

and

$$FP + GL = PS, \tag{14}$$

respectively.

(iii) The error bound (12) is minimized if the model (2) achieves least squares moment matching at $\sigma(S)$ and if the matrix $T \in \mathbb{R}^{\nu\times\nu}$ is orthogonal [3].

3. NONLINEAR SYSTEMS

Consider a system described by the equations [4]

$$\dot{x} = f(x, u), \quad y = h(x), \tag{15}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, and $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}$ such that $f(0, 0) = 0$ and $h(0) = 0$, and a signal generator described by the equations

$$\dot{\omega} = s(\omega), \quad \theta = l(\omega), \tag{16}$$

in which $\omega(t) \in \Omega$ and $\theta(t) \in \mathbb{R}$, with $\Omega \subset \mathbb{R}^\nu$ a sufficiently small [5] open, connected, invariant set containing the origin, while the mappings $s : \Omega \to \Omega$ and $h : \Omega \to \mathbb{R}$ are such that $s(0) = 0$ and $l(0) = 0$, respectively. The system (15) is assumed to be minimal, *i.e.* (locally) accessible and observable [6] at the origin.

We begin by recalling the definition of moment of a nonlinear system from (Astolfi, 2010). To this end, we first introduce some basic assumptions.

(A1)* The matrix $S = \frac{\partial s}{\partial \omega}(0)$ is non-derogatory.

(A2)* The system (16) is (locally) observable at the origin.

(A3)* The partial differential equation

$$f(\pi(\omega), l(\omega)) = \frac{\partial \pi}{\partial \omega}(\omega) s(\omega), \tag{17}$$

admits a unique solution $\pi(\cdot)$, locally defined in the neighbourhood Ω of the origin, such that $\pi(0) = 0$.

Definition 4. (Astolfi, 2010) Consider system (15) and the signal generator (16). Suppose Assumptions (A1)*-(A3)* hold. The *moment of system* (15) *at* (s, l) is defined as the mapping $\mu(\cdot) = h(\pi(\cdot))$, where $\pi(\cdot)$ is the unique solution of the partial differential equation (17).

[3] A matrix $T \in \mathbb{R}^{n\times n}$ is orthogonal if $TT^\mathsf{T} = I$ (Horn and Johnson, 1994, p.84).

[4] All mappings are assumed to be smooth, *i.e.* infinitely many times differentiable, if not otherwise stated.

[5] All statements are local, although global versions can be given.

[6] See (Nijmeijer and Van der Schaft, 1990) for the notion of local accessibility and observability.

Remark 5. Assumptions (A1)*-(A3)* ensure that the moment of system (15) at (s, l) is (locally) well-defined. Note that the partial differential equation (17) admits a unique formal (power series) solution if and only if the following non-resonance condition holds (Huang, 2004, Lemma 4.13)

$$\sigma(A) \cap \sigma_k(S) = \emptyset, \quad k = 1, 2, \ldots, \tag{18}$$

in which $A = \frac{\partial f}{\partial x}(0, 0)$ and $S = \frac{\partial s}{\partial \omega}(0)$, and

$$\sigma_k(S) = \left\{ \lambda \in \mathbb{C} \,:\, \lambda = \sum_{i=1}^{\nu} \lambda_i k_i,\ k = \sum_{i=1}^{\nu} k_i \right\}, \tag{19}$$

where $\lambda_i \in \sigma(S)$ and $k_i \in \{0, 1, \ldots, k\}$. In particular, Assumption (A3)* holds if the equilibrium $x = 0$ of the system $\dot{x} = f(x, 0)$ is (locally) exponentially stable and if the signal generator (16) is periodic.

Definition 6. (Astolfi, 2010) The system

$$\dot{\xi} = \phi(\xi, v), \quad \psi = \kappa(\xi), \tag{20}$$

with $\xi(t) \in \mathbb{R}^r$, $v(t) \in \mathbb{R}$, $\psi(t) \in \mathbb{R}$, and $\phi : \mathbb{R}^r \times \mathbb{R} \to \mathbb{R}^n$, $\kappa : \mathbb{R}^r \to \mathbb{R}$ such that $\phi(0, 0) = 0$ and $\kappa(0) = 0$, is a *model of system* (15) *at* (s, l) if its moment at (s, l) is (locally) well-defined and coincides with that of system (15), *i.e.* if the partial differential equation

$$\phi(p(\omega), l(\omega)) = \frac{\partial p}{\partial \omega}(\omega) s(\omega) \tag{21}$$

possesses a unique solution $p(\cdot)$, locally defined in the neighbourhood Ω of the origin, such that $p(0) = 0$ and

$$h(\pi(\omega)) = \kappa(p(\omega)), \tag{22}$$

where $\pi(\cdot)$ is the (unique) solution of the partial differential equation (17). In this case, system (20) is said to *match the moment of system* (15) *(or to achieve moment matching) at* (s, l). Furthermore, system (20) is a *reduced order model (of system* (15)*) at* (s, l) if $r < n$.

3.1 Least squares moment matching for nonlinear systems

We are now in a position to introduce a nonlinear enhancement of the notion of least squares moment matching.

Definition 7. Consider system (15) and the signal generator (16). Let $\tau : \Omega \to \Omega$ be a diffeomorphism such that $\tau(0) = 0$. Suppose Assumptions (A1)*-(A3)* hold. The model (20) *achieves least squares moment matching at* (s, l) if the triple (ϕ, κ, p) is a formal (power series) solution of the constrained the optimization problem

$$\begin{aligned} \text{minimize} \quad & \sup_{\omega\in\Omega} \left\| \left(\frac{\partial \mu}{\partial \omega}(\tau(\omega)) - \frac{\partial \hat{\mu}}{\partial \omega}(\tau(\omega)) \right) \frac{\partial \tau}{\partial \omega}(\omega) \right\|_{2*} \\ \text{subject to} \quad & \phi(p(\omega), l(\omega)) = \frac{\partial p}{\partial \omega}(\omega) s(\omega),\ \omega \in \Omega, \\ & \sigma(F) \cap \sigma_k(S) = \emptyset,\ k = 1, 2, \ldots, \end{aligned} \tag{23}$$

in which $\mu(\cdot) = h(\pi(\cdot))$, $\hat{\mu}(\cdot) = \kappa(p(\cdot))$, $F = \frac{\partial \phi}{\partial \xi}(0,0)$ and $S = \frac{\partial s}{\partial \omega}(0)$, where $\phi(\cdot, \cdot)$, $\kappa(\cdot)$ and $p(\cdot)$ are the optimization variables, while system (15) and the signal generator (16) (and, thus, the solution $\pi(\cdot)$ of the partial differential equation (17)) are problem data.

The *model reduction problem by least squares moment matching* for nonlinear systems can be thus posed as follows.

Problem 2. Consider system (15) and the signal generator (16). Let $\tau : \Omega \to \Omega$ be a diffeomorphism such that

$\tau(0) = 0$. Suppose Assumptions (A1)*-(A3)* hold. Let $r \in \mathbb{Z}_+$, with $2r < \nu$. Find, if possible, a model (20) of order r achieves least squares moment matching at (s, l).

We now characterize least squares moment matching in terms of the steady-state behavior of the error system

$$\dot{x} = f(x, u), \quad \dot{\xi} = \phi(\xi, u), \quad e = h(x) - \kappa(\xi), \tag{24}$$

in which $x(t) \in \mathbb{R}^n$, $\xi(t) \in \mathbb{R}^r$, $u(t) \in \mathbb{R}$, and $e(t) \in \mathbb{R}$. Generalizing Theorem 3, one may show that if the steady-state output response e_{ss} of the interconnected system (16)-(24), with $u = \theta$, is (locally) well-defined and if all solutions of the signal generator (9) are periodic, then achieving least squares moment matching corresponds to minimizing an upper bound of the *worst case r.m.s. gain* of the error system (24) with respect to the family of signals produced by the signal generator (9), defined as [7]

$$\gamma_{rms} = \sup_{\omega(\cdot) \in \mathcal{W}} \frac{\|e_{ss}\|_{rms}}{\|\omega\|_{rms}}, \tag{25}$$

where $\mathcal{W}$ is the family of signals produced by (16) with non-zero r.m.s. value.

Theorem 8. Consider system (15), the signal generator (16) and the model (20). Let $\tau : \Omega \to \Omega$ be a diffeomorphism such that $\tau(0) = 0$. Suppose Assumptions (A1)*-(A3)* hold. Assume that the zero equilibrium of the system $\dot{x} = f(x, 0)$, $\dot{\xi} = \phi(\xi, 0)$ is locally exponentially stable, that all solutions of system (16) are periodic, and that the pair $(s, \omega(0))$ is exciting [8]. Then the following statements hold.

(i) The steady-state output response of the interconnected system (16)-(24), with $u = \theta$, is (locally) well-defined and uniquely determined by the moment of the error system (24) at (s, l).

(ii) The worst case r.m.s. gain of the error system (24) with respect to the family of signals produced by the signal generator (9) is well-defined and such that

$$\gamma_{rms} \leq \sup_{\omega \in \Omega} \left\| \frac{\partial \mu}{\partial \omega}(\omega) - \frac{\partial \hat{\mu}}{\partial \omega}(\omega) \right\|_{2*}, \tag{26}$$

where $\mu(\cdot)$ and $\hat{\mu}(\cdot)$ are the moments of systems (15) and (20) at (s, l), respectively.

(iii) The error bound (26) is minimized if the model (20) achieves least squares moment matching at (s, l) and if the mapping $\tau(\cdot)$ is an isometry [9] (on Ω).

4. CONCLUSION

The model reduction problem by least squares moment matching has been studied. A recent time-domain characterization of least squares moment matching has been used to define a notion of least squares moment matching for nonlinear systems. The results presented in this work can be used to obtain new families of models achieving least squares moment matching both for linear and nonlinear systems. The natural geometric and system-theoretic interpretations of such families will be discussed in detail in a future publication.

[7] See (Isidori and Astolfi, 1992) for closely the related notion of H_∞ gain of a nonlinear system.
[8] See (Padoan et al., 2017) for the definition.
[9] See (Lee, 2013, p.332) for the definition.

REFERENCES

Aguirre, L.A. (1992). The least squares Padé method for model reduction. *Int. J. Syst. Sci.*, 23(10), 1559–1570.

Antoulas, A.C. (2005). *Approximation of large-scale dynamical systems*. SIAM, Philadelphia, PA, USA.

Astolfi, A. (2010). Model reduction by moment matching for linear and nonlinear systems. *IEEE Trans. Autom. Control*, 55(10), 2321–2336.

Boyd, S.P. and Barratt, C.H. (1991). *Linear controller design: limits of performance*. Prentice Hall, Englewood Cliffs, NJ, USA.

Boyd, S.P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge Univ. Press, Cambridge, U.K.

Gallivan, K.A., Vandendorpe, A., and Van Dooren, P. (2004). Sylvester equations and projection-based model reduction. *J. Comp. Applied Math.*, 162(1), 213–229.

Gallivan, K.A., Vandendorpe, A., and Van Dooren, P. (2006). Model reduction and the solution of Sylvester equations. In *Proc. 17th Math. Symp. Netw. Syst.*. Kyoto, Japan.

Grimme, E.J. (1997). *Krylov projection methods for model reduction*. Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

Gugercin, S. and Antoulas, A.C. (2006). Model reduction of large-scale systems by least squares. *Lin. Alg. Appl.*, 415(2-3), 290–321.

Horn, R.A. and Johnson, C.R. (1994). *Matrix analysis (2nd Ed.)*. Cambridge University Press, Cambridge, U.K.

Huang, J. (2004). *Nonlinear output regulation: theory and applications*. SIAM, Philadelphia, PA, USA.

Isidori, A. (1995). *Nonlinear control systems (3rd Ed.)*. Springer-Verlag, New York, NY, USA.

Isidori, A. and Astolfi, A. (1992). Disturbance attenuation and H_∞-control via measurement feedback in nonlinear systems. *IEEE Trans. Autom. Control*, 37(9), 1283–1293.

Lee, J.M. (2013). *Introduction to Smooth Manifolds (2nd edition)*. Springer, New York, NY, USA.

Mayo, A.J. and Antoulas, A.C. (2007). A framework for the solution of the generalized realization problem. *Lin. Alg. Appl.*, 425(2-3), 634–662.

Nakatsukasa, Y., Sète, O., and Trefethen, L.N. (2018). The AAA algorithm for rational approximation. *SIAM J. Sci. Comp.*, 40(3), A1494–A1522.

Nijmeijer, H. and Van der Schaft, A. (1990). *Nonlinear dynamical control systems*. Springer-Verlag, New York, NY, USA.

Padoan, A. (2021). On model reduction by least squares moment matching. In *Proc. 60th Conf. Decision Control*, 6901–6907. Austin, TX, USA.

Padoan, A. and Astolfi, A. (2019). Singularities and moments of nonlinear systems. *IEEE Trans. Autom. Control*, 65(8), 3647–3654.

Padoan, A., Scarciotti, G., and Astolfi, A. (2017). A geometric characterisation of the persistence of excitation condition for the solutions of autonomous systems. *IEEE Trans. Autom. Control*, 62(11), 5666 –5677.

Scarciotti, G. and Astolfi, A. (2016). Model reduction of neutral linear and nonlinear time-invariant time-delay systems with discrete and distributed delays. *IEEE Trans. Autom. Control*, 61(6), 1438–1451.

Smith, I.D. and Lucas, T.N. (1995). Least-squares moment matching reduction methods. *Electron. Lett.*, 31(11), 929–930.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Data-Driven Model Reduction by Moment Matching: One-Shot Moment Approximation through a Swapped Interconnection

Junyu Mao * **Giordano Scarciotti** *

* *Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, London, U.K.*
(e-mail: junyu.mao18@imperial.ac.uk, g.scarciotti@imperial.ac.uk)

Abstract: In this extended abstract, we present a time-domain data-driven technique for model reduction by moment matching of linear systems. We propose an algorithm, based on the so-called swapped interconnection, that (asymptotically) approximates an arbitrary number of moments of the system from a single time-domain sample. A family of reduced-order models that match the estimated moments is derived. Finally, the use of the proposed algorithm is demonstrated on the problem of model reduction of an atmospheric storm track model.

Keywords: Model reduction, system identification, moment matching, data-driven, time-domain.

1. INTRODUCTION

In a wide range of modern engineering problems, where large-scale systems are under consideration, high-order mathematical models are commonly constructed to describe the dynamics of those systems for the model-based analysis, control and prediction. However, in practice, the high dimensionality of these models poses a considerable challenge to nowadays computational power, in spite of its rapid advancement. To ease this computational burden, the problem of *model reduction* aims at reducing the complexity (*e.g.* dimensionality) of dynamical models. Informally, model reduction can be considered as the problem of approximating the important behaviours (*i.e.* input-output mapping) of a certain model by a simplified description, *e.g.*, a lower-order model. For linear systems, model reduction techniques have been extensively studied for decades, see, *e.g.*, Adamjan et al. (1971); Moore (1981); Meyer (1990). A popular family of methods is based on the interpolation framework and Krylov projection theory, see, *e.g.*, Kimura (1986); Georgiou (1999); Byrnes et al. (2001). This class is also commonly referred to as *moment matching* methods since the resulting reduced-order models match exactly the "moments" of the original system at specific frequencies.

Recently, to largely obviate the conventional need of a state-space model of the system to be reduced, some data-driven model reduction methods have been proposed in the interpolation-based moment matching framework. Among these works, we mention the seminal work of Mayo and Antoulas (2007) which introduced the so-called Loewner framework, which leverages frequency-domain samples for model reduction by moment matching. This extended abstract is based on a different framework introduced by (*e.g.* Scarciotti and Astolfi (2017)) which focuses on time-domain measurements to construct families of reduced-order models which asymptotically match the moments over time, for both linear and nonlinear systems. The experimental formulation of that approach is based on output measurements of an interconnection where the signal generator drives the system to be reduced, which we refer to as the "*direct*" interconnection in the rest of this paper.

In this work, we study a direct counterpart of Scarciotti and Astolfi (2017) based on a "*swapped*" interconnection – the system output drives a (generalized) signal generator. We show that the method proposed in this paper has some surprising advantages over Scarciotti and Astolfi (2017) in terms of sample efficiency. Also, we note that the results of this paper provide a necessary preliminary step for the development of data-driven model reduction techniques for nonlinear systems.

The remainder of this paper is organized as follows. In Section 2.1 we recall the theory of moment matching. Section 2.2 presents a data-driven approach that estimates the moments asymptotically, together with the associated reduced-order models. Finally, in Section 2.3, the proposed approach is demonstrated by a storm track model.

Notation $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real numbers and complex numbers respectively. $\mathbb{R}_{\geq 0}$ ($\mathbb{R}_{>0}$) denotes the set of non-negative (positive) real numbers. $\mathbb{C}_0$ ($\mathbb{C}_{<0}$) denotes the set of complex numbers with zero (negative) real part. The set of non-negative integers is denoted by $\mathbb{Z}_{\geq 0}$. The identity matrix is denoted by the symbol I, and $\sigma(A)$ denotes the spectrum of a square matrix A. The symbol $\otimes$ represents the Kronecker product. The operator $\mathrm{vec}(A)$ indicates the vectorization of a matrix $A \in \mathbb{R}^{n\times m}$. $A^\top$ denotes the transpose of any matrix A. ι denotes the imaginary unit.

2. DATA-DRIVEN MOMENT MATCHING BY SWAPPED INTERCONNECTION

2.1 Moment Matching via System Interconnections – Recalled

In this section we recall the notion of moment for linear systems and its relation to the steady-state of certain interconnections. First consider a linear, single-input, single-output (SISO) [1],

[1] The results can be easily extended to systems with multiple inputs and multiple outputs.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

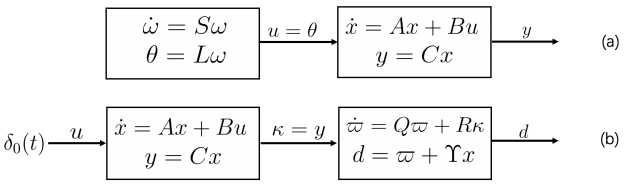


Fig. 1. Diagrammatic illustrations of the *direct* interconnection (a) and the *swapped* interconnection (b).

continuous-time system, described by equations of the form

$$\dot x = Ax + Bu, \qquad y = Cx, \tag{1}$$

with state $x(t) \in \mathbb{R}^n$, input $u(t) \in \mathbb{R}$, output $y(t) \in \mathbb{R}$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$ and $C \in \mathbb{R}^{1\times n}$. The associated transfer function is $W(s) = C(sI - A)^{-1}B$. Assume system (1) is minimal, *i.e.*, both controllable and observable. In the following we define the moments of system (1).

Definition 1. *The 0-moment of system (1) at $s_i \in \mathbb{C}\backslash\sigma(A)$ is the complex number $\eta_0(s_i) = W(s_i)$. The k-moment of system (1) at s_i is the complex number $\eta_k(s_i) = \frac{(-1)^k}{k!}\left[\frac{d^k}{ds^k}W(s)\right]_{s=s_i}$, where $k \geq 1$ is an integer.*

Definition 1 is a direct characterization of moments based on the transfer function and consequently it can only be used for linear, time-invariant systems. To obviate this restriction, Astolfi (2010a) noted that the moments are in a one-to-one relation with the (well-defined) steady-state responses of interconnections between the system and some "signal generators". Consider the signal generator

$$\dot\omega = S\omega, \qquad \theta = L\omega, \tag{2}$$

with $\omega(t) \in \mathbb{R}^\nu$, $u(t) \in \mathbb{R}$, $S \in \mathbb{R}^{\nu\times\nu}$ and $L \in \mathbb{R}^{1\times\nu}$, and the interconnection between this generator and system (1), namely

$$\dot\omega = S\omega, \qquad \dot x = Ax + BL\omega, \qquad y = Cx. \tag{3}$$

This interconnection is schematically represented in Fig. 1(a) and is called *direct* interconnection. Likewise, consider the signal generator

$$\dot\varpi = Q\varpi + R\kappa, \qquad d = \varpi + \Upsilon x, \tag{4}$$

with $\varpi(t) \in \mathbb{R}^\nu$, $\kappa(t) \in \mathbb{R}$, $d(t) \in \mathbb{R}^\nu$, $Q \in \mathbb{R}^{\nu\times\nu}$, $R \in \mathbb{R}^{\nu\times 1}$ and $\Upsilon \in \mathbb{R}^{\nu\times\nu}$, and the interconnection between this generator and system (1), namely

$$\dot x = Ax + Bu, \qquad \dot\varpi = Q\varpi + RCx, \qquad d = \varpi + \Upsilon x. \tag{5}$$

This interconnection is schematically represented in Fig. 1(b) with $u = \delta_0$, where δ_0 indicates the Dirac-delta, and is called *swapped* interconnection.

To streamline the presentation, we now introduce a series of assumptions.

Assumption 1. *For the signal generator (2), the pair (S, L) is observable. For the signal generator (4), the pair (Q, R) is controllable.*

Assumption 2. *System (1) is asymptotically stable, i.e., $\sigma(A) \subset \mathbb{C}_{<0}$. The matrix S has simple eigenvalues with $\sigma(S) \subset \mathbb{C}_0$. The pair (S, ω_0) is excitable (see Padoan et al. (2016)).*

Assumption 3. *System (1) is asymptotically stable. The matrix Q has simple eigenvalues with $\sigma(Q) \subset \mathbb{C}_0$. The initial condition $x(0)$ is 0.*

Note that under the asymptotic stability of system (1), the assumption that the initial condition of the system is zero is without loss of generality.

We now recall a result that connects moments with the steady states of interconnections (3) and (5).

Theorem 1 (see Astolfi (2010b))**.** *Let $s_i \in \mathbb{C} \setminus \sigma(A)$, for $i = 1, \dots, \rho$. Consider system (1) and assume that for the signal generators (2) and (4), the matrices S and Q are non-derogatory* [2] *with characteristic polynomial $p(s) = \prod_{i=1}^{\rho}(s - s_i)^{k_i}$, and $\nu = \sum_{i=1}^{\rho} k_i$. Then, the moments of system (1), namely $\eta_0(s_1), \dots, \eta_{k_1-1}(s_1), \dots, \eta_0(s_\rho), \dots, \eta_{k_\rho-1}(s_\rho)$, are in a one-to-one relation to*

- *the matrix $C\Pi$, in which $\Pi \in \mathbb{R}^{n\times\nu}$ is the unique solution of the Sylvester equation*
$$\Pi S = A\Pi + BL, \tag{6}$$
provided Assumption 1 holds.
- *the steady state of the output $y(t)$ of the interconnection (3), provided that Assumptions 1 and 2 hold.*
- *the matrix ΥB, in which $\Upsilon \in \mathbb{R}^{\nu\times n}$ is the unique solution of the Sylvester equation*
$$Q\Upsilon = \Upsilon A + RC, \tag{7}$$
provided Assumption 1 holds.
- *the steady state of the output $d(t)$ of the interconnection (5), provided that $u = \delta_0$ and Assumptions 1 and 3 hold.*

Given a set of moments, Theorem 1 provides four alternative characterizations of those moments. In particular, the fact that moments are in one-to-one relation with certain steady states opens the possibility of computing them from time-domain measurements of these signals.

Throughout this paper, we focus on the swapped interconnection. We conclude this section, by recalling a family of reduced-order models that achieve moment matching at (Q, R), *i.e.*, the reduced-order models have the same moments at the frequencies $s_i \in \sigma(Q)$. This family is described by the equations

$$\dot\xi = (Q - RH)\xi + \Upsilon Bu, \qquad \psi = H\xi, \tag{8}$$

for any $H \in \mathbb{R}^{1\times\nu}$ such that $\sigma(Q - RH) \cap \sigma(Q) = \emptyset$. This family contains all the ν-order models which achieve moment matching at $\sigma(Q)$. In this paper we do not need the family of reduced-order models associated to (S, L), hence, it is omitted.

2.2 On-Line Moment Matching from Experimental Data

In this section we develop a data-driven approach for (asymptotically) estimating the moments (more precisely ΥB) of the system, purely based on the measurements of the state $\varpi(t)$ of the user-defined signal generator. We first state an instrumental observation about the signal $d(t)$.

Remark 1. *Let Υ be the unique solution of equation (7). Note that the signal $d(t)$ can be expressed as the output of the system described by*

$$\dot d = Qd + \Upsilon Bu, \qquad \chi = d. \tag{9}$$

This follows from interconnection (5) and equation (7), namely

$$\begin{aligned}\dot d &= \dot{\varpi} + \Upsilon\dot x = Q\varpi + Ry + \Upsilon\left(Ax + Bu\right)\\ &= Q\varpi + \underbrace{(RC + \Upsilon A)}_{Q\Upsilon}x + \Upsilon Bu = Q(\varpi + \Upsilon x) + \Upsilon Bu\\ &= Qd + \Upsilon Bu.\end{aligned}$$

We are now ready to explicitly describe the dynamics of the signal $d(t)$ for the interconnection (5).

[2] A matrix is non-derogatory if its characteristic and minimal polynomial coincide.

Proposition 1. *Consider the interconnection (5) with $\varpi(0) = 0$ and $u = \delta_0$. Suppose Assumption 3 holds. Then, for all $t \in \mathbb{R}_{\geq 0}$, the response of system (9) is*

$$d(t) = e^{Qt}\Upsilon B. \tag{10}$$

Proof. Under the assumption that $x(0) = 0$ and $\varpi(0) = 0$, we have that $d(0) = 0$. The result follows trivally by computing the impulse response of the system (9). □

Exploiting Proposition 1, we can characterize the relation between the state of system (1) and the state of the signal generator (4), through the matrix Υ, namely

$$\varpi(t) + \Upsilon x(t) = d(t) = e^{Qt}\Upsilon B, \tag{11}$$

which underpins the following theorem.

Theorem 2. *Consider the interconnection (5) with $\varpi(0) = 0$ and $u = \delta_0$. Suppose Assumptions 1 and 3 hold. Then, ΥB characterizes the steady state $\varpi_{ss}(t)$ of $\varpi(t)$*

$$\varpi_{ss}(t) = e^{Qt}\Upsilon B. \tag{12}$$

Proof. As system (1) is asymptotically stable (by Assumption 3), the impulse response of $x(t)$ will exponentially decay to zero as the time goes to infinity, *i.e.*, $\lim_{t\to+\infty} x(t) = 0$. By Assumption 3, Q has simple eigenvalues with $\sigma(Q) \in \mathbb{C}_0$. Consequently, $\varpi_{ss}(t)$ is well defined for all times and e^{Qt} is a signal that is persistent in time. Thus, equation (11) at steady state reduces to (12).

□

We are ready to present the estimation in the following theorem.

Theorem 3. *Consider the interconnection (5) with $\varpi(0) = 0$ and $u = \delta_0$. Suppose Assumptions 1 and 3 hold. Then,*

$$\widetilde{\Upsilon B}_k := e^{-Qt_k}\varpi(t_k) \tag{13}$$

is an online estimate of ΥB with the following asymptotic property: there exist a sequence $\{t_k\}$ such that

$$\lim_{k\to+\infty} \widetilde{\Upsilon B}_k = \Upsilon B. \tag{14}$$

Proof. Multiplying e^{-Qt} to both sides of equation (11), evaluated at t_k, yields $\widetilde{\Upsilon B}_k + \epsilon(t_k) = \Upsilon B$, in which $\epsilon(t_k) = e^{-Qt_k}\Upsilon x(t_k) = e^{-Qt_k}\Upsilon e^{At_k}B$ is an exponentially decaying error signal. It follows by Theorem 2 that

$$\lim_{k\to+\infty}\left(\widetilde{\Upsilon B}_k - \Upsilon B\right) = \lim_{t_k\to+\infty} \epsilon(t_k) = 0, \tag{15}$$

as long as the time sequence $\{t_k\}$ indexed by k grows unbounded to $+\infty$. □

We summarize the above results in Algorithm 1.

Remark 2. *In the original model-based method proposed in Astolfi (2010b), the computation of the matrix Υ involves solving the Sylvester equation (7), which has a computational complexity*[3] *of order $\mathcal{O}(n^3)$. In comparison, Algorithm 1 only involves a matrix exponential and a matrix multiplication of order ν, hence both have a computational complexity of roughly $\mathcal{O}(\nu^3)$ (Al-Mohy and Higham (2010)). Thus, the proposed method is computationally efficient than the original method as $\nu \ll n$.*

Algorithm 1 On-Line Approximation of ΥB.

Input: a *sufficiently* small tolerance $\eta_{\Upsilon B} > 0$, $k = 0$
while 1 **do**
  Obtain data $\varpi(t_k)$ and record time instant t_k
  Compute $\widetilde{\Upsilon B}_k$

$$\widetilde{\Upsilon B}_k = e^{-Qt_k}\varpi(t_k)$$

  if $\left\|\widetilde{\Upsilon B}_k - \widetilde{\Upsilon B}_{k-1}\right\| \leq \frac{\eta_{\Upsilon B}}{t_k - t_{k-1}}$ **then**
    Break
  end if
  $k = k + 1$
end while
Return: $\widetilde{\Upsilon B}_k$

Remark 3. *A surprising advantage of the proposed approach based on the swapped interconnection is the sample efficiency, i.e., the number of samples required to properly estimate the moments. In Scarciotti and Astolfi (2017), the method, which is based on a direct interconnection, requires data from at least ν time instants to estimate ν moments. In contrast, within Algorithm 1, data from only one time instant suffices to fully estimate an arbitrarily large number of moments.*

More importantly, the data-driven approach under the swapped interconnection plays a core role in paving the path for developing a data-driven framework for the two-sided moment matching (Ionescu (2015)).

With the estimation of moments from Algorithm 1, we are able to construct the associated reduced-order models by replacing the moments ΥB with its online estimate $\widetilde{\Upsilon B}_k$. For all $t_k > 0$, the family of reduced-order models for system (1) at the pair $(\sigma(Q), t_k)$ is defined as

$$\dot{\xi} = (Q - RH_k)\xi + \widetilde{\Upsilon B}_k u, \qquad \psi = H_k\xi, \tag{16}$$

where $H_k \in \mathbb{R}^{1\times\nu}$ is a free parametrization as long as $\sigma(Q - RH_k) \cap \sigma(Q) = \emptyset$. Furthermore, Astolfi (2010a) shows that the free matrix H_k can be chosen to enforce additional properties of the reduced-order models, *e.g.*, stability, passivity, prescribed relative degree, spectrum and zeros.

2.3 Numerical Example

In this section we illustrate the use of Algorithm 1 with the *Eady Example*[4], *i.e.*, an atmospheric storm track model which is a widely-used benchmark (Antoulas et al. (2000)) for a variety of model reduction techniques, see *e.g.*, Antoulas (2005). In this work, we use the interpolation points $\pm 0.1\iota, \pm 0.58\iota, \pm 0.8\iota, \pm 0.9\iota, \pm 1.08\iota, \pm 1.26\iota, \pm 2.5\iota, \pm 7.0\iota$, which are mostly the major peaks of the bode plot of the storm track system. This selection results in a reduced order of 16. The determination of the free parametrization H_k is based on assigning the eigenvalues of F_k (see Astolfi (2010a)).

Fig. 2 shows the bode plots of the full order system (solid/blue line) and the reduced-order model (dashed/red line) constructed using the estimated moments $\widetilde{\Upsilon B}_k$ obtained at the end of an online experiment of 18 seconds. It shows that the moments between the two systems are matched at those interpolated frequencies (green circles). In addition, the time history of

[3] This is the computational cost for general dense A and B, *e.g.,* using the classic Bartels–Stewart algorithm. If A and B are very sparse, the computational cost can be further reduced to $\mathcal{O}(n^2 r)$ or $\mathcal{O}(nr^2)$ by ADI method (Benner et al. (2009)), where r is the dimension of the low-rank factor.

[4] The data can be downloaded at `http://slicot.org/20-site/126-benchmark-examples-for-model-reduction`.

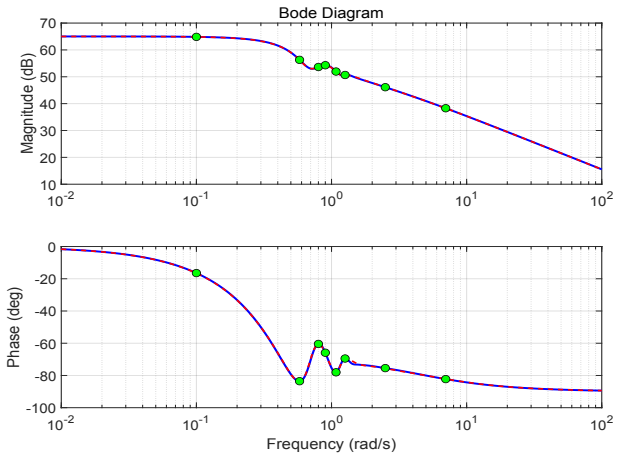


Fig. 2. Bode plots of the storm track model (solid/blue line) and of the data-driven reduced-order model (dashed/red line) constructed with the approximation $\widehat{\Upsilon B}_k$ for $t_k = 18$ s. The circles represent the set of interpolation points.

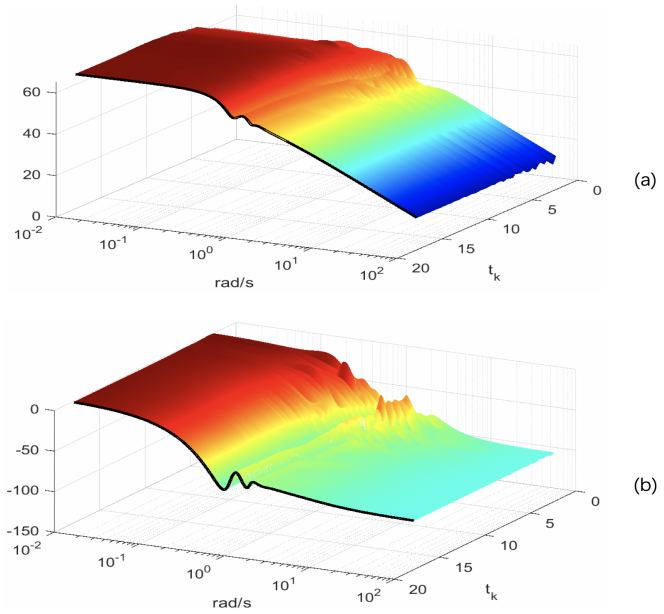


Fig. 3. The surface represents the magnitude (a) and phase (b) of the transfer function of the reduced-order model as a function of t_k, with $2.44 \leq t_k \leq 18$ s. The solid/black line indicates the transfer function of the reduced-order model constructed using the exact moments ΥB.

the bode plot of the reduced-order models obtained within an experimental time interval $\mathcal{T} = \{t_k \in \mathbb{R} : 2.44 \leq t_k \leq 18\}$ (of unit *second*) is shown in Fig. 3. The solid/black line represents the model obtained with the exact ΥB. The plots illustrate the evolution of the data-driven reduced-order model over time, demonstrating that it converges to the exact moment-matching reduced-order model as the time goes to infinity.

3. CONCLUSIONS

We have presented a theoretical framework and a time-domain data-driven method to solve the model reduction problem by moment matching, without knowing the state-space model of the linear system to be reduced. Firstly, an algorithm under the swapped interconnection has been proposed to asymptotically approximate the moments of an unknown system. Then, by exploiting the obtained approximations, a family of (parametrized) reduced-order models has been given. Finally, we have illustrated the performances of the proposed algorithm by means of a widely-used benchmark.

REFERENCES

Adamjan, V.M., Arov, D.Z., and Kren, M. (1971). Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem. *Mathematics of the USSR-Sbornik*, 15(1), 31.

Al-Mohy, A.H. and Higham, N.J. (2010). A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3), 970–989.

Antoulas, A.C. (2005). *Approximation of large-scale dynamical systems*. SIAM.

Antoulas, A.C., Sorensen, D.C., and Gugercin, S. (2000). A survey of model reduction methods for large-scale systems. Technical report.

Astolfi, A. (2010a). Model reduction by moment matching for linear and nonlinear systems. *IEEE Transactions on Automatic Control*, 55(10), 2321–2336.

Astolfi, A. (2010b). Model reduction by moment matching, steady-state response and projections. In *Proceedings of the 49th IEEE Conference on Decision and Control*, 5344–5349. IEEE.

Benner, P., Li, R.C., and Truhar, N. (2009). On the ADI method for Sylvester equations. *Journal of Computational and Applied Mathematics*, 233(4), 1035–1045.

Byrnes, C.I., Georgiou, T.T., and Lindquist, A. (2001). A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint. *IEEE Transactions on Automatic Control*, 46(6), 822–839.

Georgiou, T.T. (1999). The interpolation problem with a degree constraint. *IEEE Transactions on Automatic Control*, 44(3), 631–635.

Ionescu, T.C. (2015). Two-sided time-domain moment matching for linear systems. *IEEE Transactions on Automatic Control*, 61(9), 2632–2637.

Kimura, H. (1986). Positive partial realization of covariance sequences. *Modeling, identification and robust control*, 499–513.

Mayo, A. and Antoulas, A. (2007). A framework for the solution of the generalized realization problem. *Linear algebra and its applications*, 425(2-3), 634–662.

Meyer, D.G. (1990). Fractional balanced reduction: Model reduction via fractional representation. *IEEE Transactions on Automatic Control*, 35(12), 1341–1345.

Moore, B. (1981). Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1), 17–32.

Padoan, A., Scarciotti, G., and Astolfi, A. (2016). A geometric characterisation of persistently exciting signals generated by continuous-time autonomous systems. *IFAC-PapersOnLine*, 49(18), 826–831.

Scarciotti, G. and Astolfi, A. (2017). Data-driven model reduction by moment matching for linear and nonlinear systems. *Automatica*, 79, 340–351.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Neural Network Optimal Feedback Control with Guaranteed Local Stability ⋆

Tenavi Nakamura-Zimmerer *,** **Qi Gong** *

* *Applied Mathematics Department, University of California, Santa Cruz, CA 95064, USA (e-mail: tenakamu@ucsc.edu, qgong@ucsc.edu)*
** *Flight Dynamics Branch, NASA Langley Research Center, Hampton, VA 23666, USA.*

Abstract: Recent work has demonstrated the potential of applying supervised learning to train neural networks which approximate optimal feedback laws. In this talk, we show that some neural networks with good test accuracy can fail to even locally stabilize the dynamics. To address this challenge, we propose some novel neural network architectures which guarantee local asymptotic stability while still closely approximating optimal feedback laws on large domains.

Keywords: Large-scale optimal control problems, numerical methods for optimal control, learning for control, machine learning, stability of nonlinear systems.

1. INTRODUCTION

Designing optimal feedback controllers for high-dimensional nonlinear systems remains an outstanding challenge. Even when the system dynamics are known, one needs to solve a Hamilton-Jacobi-Bellman (HJB) partial differential equation (PDE) in as many variables as there are states, leading to the well-known curse of dimensionality.

Recent work has demonstrated the promise of supervised learning with neural networks (NNs) as one potential approach for handling challenging, high-dimensional problems. The main idea is to fit a model to data generated by solving many open loop optimal control problems (OCPs), thus obtaining an approximate optimal feedback controller. Further details can be found in e.g. Kang and Wilcox (2017), Izzo and Öztürk (2021), Nakamura-Zimmerer et al. (2021a,b), and Azmi et al. (2021).

Despite progress in the methodology, much less work has been done to study and improve the stability and reliability of these NN controllers. To see why this is needed, if we train a set of NNs to control a Burgers'-type PDE (12), a surprisingly large number of these fail to stabilize the system despite having good test accuracy. Figure 1 shows a simulation with one such NN controller.

Tailor and Izzo (2019) have also pointed out that test accuracy incompletely characterizes the performance of NN controllers, and suggest some more practical evaluations of optimality and stability. Izzo et al. (2021) study linear stability near a desired equilibrium, linear time delay stability, and stability around a nominal trajectory.

The purpose of this talk is to bring attention to stability issues with NN-controlled systems and to discuss recent work by Nakamura-Zimmerer et al. (2021b, 2022a,b) exploring NN architectures which can potentially mitigate these challenges. Some of these architectures can guarantee, at a minimum, local asymptotic stability (LAS) of the system. At the same time, these NNs can still learn the optimal feedback law and thus provide semi-global stability and optimality. We compare the new architectures to standard NNs by means of several practical closed loop stability and optimality tests. As a testbed we use the Burgers'-type PDE system (12), which is nonlinear, open loop unstable, and high-dimensional.

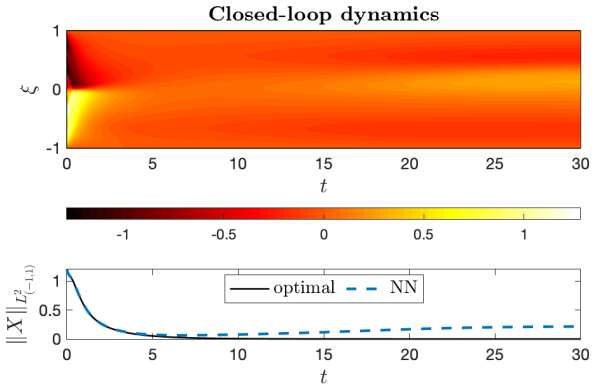


Fig. 1. Closed loop simulation of the Burgers'-type PDE (12) showing instability. Feedback control is based on an NN approximation of the solution of the HJB PDE; see Nakamura-Zimmerer et al. (2021a). The top plot shows the state $X(t, \xi)$, where ξ is the spatial variable.

2. PROBLEM SETTING

We focus on infinite horizon nonlinear OCPs of the form

$$\begin{cases} \underset{\boldsymbol{u}(\cdot)\in\mathbb{U}}{\text{minimize}} \;\; J\left[\boldsymbol{u}(\cdot)\right] = \int_0^\infty \mathcal{L}(\boldsymbol{x},\boldsymbol{u})dt, \\ \text{subject to } \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u}), \\ \qquad\qquad \boldsymbol{x}(0) = \boldsymbol{x}_0. \end{cases} \tag{1}$$

⋆ This work was supported with funding from the Air Force Office of Scientific Research (AFOSR) under grant FA9550-21-1-0113, the National Science Foundation (NSF) under grant no. 2134235, and the University of California, Santa Cruz, Baskin School of Engineering Dissertiation Year Fellowship.

Here $\boldsymbol{x} : [0,\infty) \to \mathbb{R}^n$ is the state, $\boldsymbol{u} : [0,\infty) \to \mathbb{U}$ is the control with box constraints $\mathbb{U} \subseteq \mathbb{R}^m$, and $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}^n$ is a continuously differentiable ($\mathcal{C}^1$) vector field. We consider running costs $\mathcal{L} : \mathbb{R}^n \times \mathbb{U} \to [0,\infty)$ of the form

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{u}) = q\left(\boldsymbol{x}-\boldsymbol{x}_f\right) + \left(\boldsymbol{u}-\boldsymbol{u}_f\right)^T \boldsymbol{R}\left(\boldsymbol{u}-\boldsymbol{u}_f\right), \qquad (2)$$

where $(\boldsymbol{x}_f,\boldsymbol{u}_f) \in \mathbb{R}^n \times \mathbb{U}$ is a (possibly unstable) equilibrium, $\boldsymbol{R} \in \mathbb{R}^{m\times m}$ is a positive definite matrix, and $q : \mathbb{R}^n \to [0,\infty)$ is a smooth, positive semi-definite function which is zero at $\boldsymbol{x}_f$. This standard cost function is a natural choice for regularization or set-point tracking problems. We make the standard assumptions that $\boldsymbol{u}_f$ is contained in an open subset of $\mathbb{U}$ and that the OCP (1) is well-posed, i.e. that an optimal control $\boldsymbol{u}^*(t)$ exists such that $J\left[\boldsymbol{u}^*(\cdot)\right] < \infty$ and $\lim_{t\to\infty} \mathcal{L}\left(\boldsymbol{x}^*(t),\boldsymbol{u}^*(t)\right) = 0$.

Due to real-time application requirements, we would like to design a closed loop feedback controller, $\boldsymbol{u} = \boldsymbol{u}^*(\boldsymbol{x})$, which can be evaluated online given any measurement of $\boldsymbol{x}$. The mathematical framework for designing such an optimal feedback policy is the HJB equation.

Define the *value function* $V : \mathbb{R}^n \to \mathbb{R}$ as the optimal cost-to-go of (1) starting at the point $\boldsymbol{x}(0) = \boldsymbol{x}$, i.e. $V(\boldsymbol{x}) \coloneqq J\left[\boldsymbol{u}^*(\cdot)\right]$. Under appropriate conditions, the value function is the solution of the steady state HJB PDE,

$$\min_{\boldsymbol{u}\in\mathbb{U}} \mathcal{H}\left(\boldsymbol{x}, V_{\boldsymbol{x}}, \boldsymbol{u}\right) = 0, \qquad V\left(\boldsymbol{x}_f\right) = 0, \qquad (3)$$

where $V_{\boldsymbol{x}} \coloneqq \left[\partial V/\partial \boldsymbol{x}\right]^T$ and we define the Hamiltonian

$$\mathcal{H}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{u}) \coloneqq \mathcal{L}(\boldsymbol{x},\boldsymbol{u}) + \boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u}). \qquad (4)$$

If (3) can be solved (in the viscosity sense), then it provides both necessary and sufficient conditions for optimality. Furthermore, the optimal feedback control is then obtained from the the Hamiltonian minimization condition,

$$\boldsymbol{u}^*(\boldsymbol{x}) = \boldsymbol{u}^*\left(\boldsymbol{x}; V_{\boldsymbol{x}}(\boldsymbol{x})\right) = \arg\min_{\boldsymbol{u}\in\mathbb{U}} \mathcal{H}\left(\boldsymbol{x}, V_{\boldsymbol{x}}, \boldsymbol{u}\right). \qquad (5)$$

To circumvent the challenge of directly solving the HJB equation (3), we can leverage the necessary conditions for optimality, well-known in optimal control as Pontryagin's Minimum Principle (PMP):

$$\lim_{t_f\to\infty} \begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u}^*(\boldsymbol{x};\boldsymbol{\lambda})), & \boldsymbol{x}(0) = \boldsymbol{x}_0, \\ \dot{\boldsymbol{\lambda}}(t) = -\mathcal{H}_{\boldsymbol{x}}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{u}^*(\boldsymbol{x};\boldsymbol{\lambda})), & \boldsymbol{\lambda}(t_f) = \boldsymbol{0}. \end{cases} \qquad (6)$$

Here $\boldsymbol{\lambda} : [0,\infty) \to \mathbb{R}^n$ is called the *costate*. If we assume that the solution of (6) is optimal, then along the trajectory $\boldsymbol{x} = \boldsymbol{x}^*(t;\boldsymbol{x}_0)$ we have

$$\begin{cases} V(\boldsymbol{x}) = \int_t^\infty \mathcal{L}\left(\boldsymbol{x}(s),\boldsymbol{u}^*(s)\right) ds, \\ V_{\boldsymbol{x}}(\boldsymbol{x}) = \boldsymbol{\lambda}(t), \quad \boldsymbol{u}^*(\boldsymbol{x}) = \boldsymbol{u}^*(t). \end{cases} \qquad (7)$$

In supervised learning, the BVP (6) is solved for different initial conditions in the domain of interest. This yields a data set of optimal value function, gradient, and control values which can be used to train feedback controllers.

3. STABILITY-ENHANCING ARCHITECTURES FOR OPTIMAL FEEDBACK DESIGN

Our goal is to learn a feedback policy which approximates the optimal control, i.e. $\widehat{\boldsymbol{u}}(\boldsymbol{x}) \approx \boldsymbol{u}^*(\boldsymbol{x})$. Due to the complex and sometimes unpredictable behavior of NNs, there is a clear need for designing feedback controllers with *built-in* stability properties.

Previously, Nakamura-Zimmerer et al. (2021b, 2022a) proposed V-*QRnet* (originally just called *QRnet*), λ-*QRnet*, and u-*QRnet*. These controllers combine a linear quadratic regulator (LQR) with NNs. The LQR terms are good approximations of the optimal control near $\boldsymbol{x}_f$, and empirically improve stability. Meanwhile, the NNs are intended to capture nonlinearities and thereby learn the nonlinear optimal feedback over a large domain.

However, none of these architectures can *guarantee* LAS, motivating the pursuit of alternative designs. In this talk we describe one of the novel NN architectures introduced in recent work by Nakamura-Zimmerer et al. (2022b). This architecture, called, u_{Jac}-*QRnet*, guarantee LAS of $\boldsymbol{x}_f$ while retaining the ability to approximate the nonlinear optimal control semi-globally.

3.1 Novel "Jacobian" QRnet *controller*

Here we describe the stability-enhancing architectures, u-*QRnet* and u_{Jac}-*QRnet*, which combine NNs with local LQR approximations. LQR is a linear state feedback law which is computed by linearizing the nonlinear OCP (1) about $(\boldsymbol{x}_f,\boldsymbol{u}_f)$ to obtain

$$\boldsymbol{u}^{\text{LQR}}(\boldsymbol{x}) = \boldsymbol{u}_f - \boldsymbol{K}(\boldsymbol{x}-\boldsymbol{x}_f). \qquad (8)$$

Here $\boldsymbol{K} \in \mathbb{R}^{m\times n}$ is the LQR gain matrix and is computed by solving a continuous algebraic Riccati equation. We note that LQR is in fact a first order approximation of the optimal control, i.e. $\left[\partial \boldsymbol{u}^*/\partial \boldsymbol{x}\right](\boldsymbol{x}_f) = -\boldsymbol{K}$.

Let us first review u-*QRnet* from Nakamura-Zimmerer et al. (2022a). This models the full nonlinear optimal control by combining (8) with an NN, $\boldsymbol{\mathcal{N}} : \mathbb{R}^n \to \mathbb{R}^m$:

$$\widehat{\boldsymbol{u}}(\boldsymbol{x}) = \boldsymbol{u}^{\text{LQR}}(\boldsymbol{x}) + \boldsymbol{\mathcal{N}}(\boldsymbol{x}) - \boldsymbol{\mathcal{N}}(\boldsymbol{x}_f). \qquad (9)$$

To construct u_{Jac}-*QRnet*, we cancel out the linear contribution of the NN:

$$\widehat{\boldsymbol{u}}(\boldsymbol{x}) = \boldsymbol{u}^{\text{LQR}}(\boldsymbol{x}) - \left[\tfrac{\partial \boldsymbol{\mathcal{N}}}{\partial \boldsymbol{x}}(\boldsymbol{x}_f)\right](\boldsymbol{x}-\boldsymbol{x}_f) + \boldsymbol{\mathcal{N}}(\boldsymbol{x}) - \boldsymbol{\mathcal{N}}(\boldsymbol{x}_f). \qquad (10)$$

The Jacobian term is key to guaranteeing LAS by construction. Note that (9) and (10) are shown without saturation constraints, but it is easy to smoothly incorporate these directly into the model; see Nakamura-Zimmerer et al. (2022b) for details. Nakamura-Zimmerer et al. (2022b) also propose a number of other models which are not covered here.

A drawback of (10) is that the Jacobian $\left[\partial \boldsymbol{\mathcal{N}}/\partial \boldsymbol{x}\right](\boldsymbol{x}_f)$ must be evaluated during each forward pass during training. This makes training the model more expensive than u-*QRnet* which does not include this term. *After training, however, we can store the Jacobian matrix in memory so that it does not have to be recomputed online.* Therefore online evaluation is just as fast as u-*QRnet*.

3.2 Local stability guarantees and approximation capacity

Like u-*QRnet* (9), the new u_{Jac}-*QRnet* (10) automatically makes the goal state $\boldsymbol{x}_f$ an equilibrium [1]. Moreover, if we linearize the feedback control $\widehat{\boldsymbol{u}}(\cdot)$ at $\boldsymbol{x}_f$ then we recover the LQR control gain (8). This property is desirable because LQR locally asymptotically stabilizes $\boldsymbol{x}_f$, and hence the proposed controllers provide LAS by construction. This is stated formally in Proposition 1, whose proof is straightforward.

[1] Note that V-*QRnet* does *not* always make $\boldsymbol{x}_f$ an equilibrium.

Proposition 1. (LAS guarantee). Suppose $\widehat{\boldsymbol{u}}(\cdot)$ is a feedback policy specified by (10). Then $[\partial\widehat{\boldsymbol{u}}/\partial\boldsymbol{x}](\boldsymbol{x}_f) = -\boldsymbol{K}$ and $\boldsymbol{x}_f$ is an LAS equilibrium of the NN-controlled system, $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \widehat{\boldsymbol{u}}(\boldsymbol{x}))$.

LAS is a critical but bare minimum requirement. To achieve the ultimate goal of semi-global stability and optimality through NN training, (10) must also be able to approximate $\boldsymbol{u}^*(\cdot)$ with sufficient accuracy. Because $\boldsymbol{u}^*(\cdot)$ is in general not everywhere analytic, we cannot directly use the Taylor series-like structure of (10) to show this is possible. Nevertheless, for OCPs like (1) we expect $\boldsymbol{u}^*(\cdot)$ to be at least continuous and locally $\mathcal{C}^1$. Then we can apply Theorem 2 below, which shows that NNs of the form (10) are universal approximators for such functions.

Theorem 2. (Universal approximation). Let $\boldsymbol{x}_f = \boldsymbol{0}$ be an interior point of the compact set $\mathbb{X} \subset \mathbb{R}^n$. Suppose $\boldsymbol{u} \in \mathcal{C}(\mathbb{X} \to \mathbb{R}^m)$, $\boldsymbol{u}(\boldsymbol{0}) = \boldsymbol{0}$, and $\boldsymbol{u}(\cdot)$ is $\mathcal{C}^1$ in a neighborhood of $\boldsymbol{0}$. Then for all $\varepsilon > 0$, there exists a feedforward NN with $\mathcal{C}^1$ bounded, non-constant activation functions, $\boldsymbol{\mathcal{N}} \in \mathcal{C}^1(\mathbb{X} \to \mathbb{R}^m)$, such that for all $\boldsymbol{x} \in \mathbb{X}$,

$$\left\| \boldsymbol{u}(\boldsymbol{x}) - \left(\left[\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}(\boldsymbol{0}) - \frac{\partial \boldsymbol{\mathcal{N}}}{\partial \boldsymbol{x}}(\boldsymbol{0}) \right] \boldsymbol{x} + \boldsymbol{\mathcal{N}}(\boldsymbol{x}) - \boldsymbol{\mathcal{N}}(\boldsymbol{0}) \right) \right\|_1 < \varepsilon. \tag{11}$$

The proof applies the Stone-Weierstrass theorem and a universal approximation theorem from Hornik (1991). Details are given in Nakamura-Zimmerer et al. (2022b).

4. NUMERICAL RESULTS

Here we compare the proposed controllers to standard feedforward NNs trained to approximate the value function, its gradient, and the optimal control. We refer to these as V-NN, λ-NN, and u-NN, respectively. We also compare to *V-QRnet*, *λ-QRnet*, and *u-QRnet*. The results indicate that the new architectures are indeed able to accurately learn the optimal feedback control while guaranteeing local stability.

As a testbed we revisit the Burgers' stabilization OCP from Nakamura-Zimmerer et al. (2021b). This is a high-dimensional nonlinear OCP formulated by pseudospectral discretization of an unstable version of a Burgers' PDE. Briefly, the problem can be summarized as

$$\begin{cases} \min\limits_{\boldsymbol{u}(\cdot)} \; J[\boldsymbol{u}(\cdot)] = \int_0^\infty \left(\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{u}^T \boldsymbol{R} \boldsymbol{u} \right) dt, \\ \text{s.t.} \;\; \dot{\boldsymbol{x}} = -\frac{1}{2} \boldsymbol{D}\boldsymbol{x} \circ \boldsymbol{x} + \nu \boldsymbol{D}^2 \boldsymbol{x} + \boldsymbol{\alpha} \circ \boldsymbol{x} \circ e^{-\beta \boldsymbol{x}} + \boldsymbol{B}\boldsymbol{u}. \end{cases} \tag{12}$$

Here $\boldsymbol{x} : [0, \infty) \to \mathbb{R}^n$ represents the PDE state $X(t, \xi)$ collocated at spatial coordinates $\xi_j = \cos(j\pi/n)$, $j = 1, \ldots, n$, $\boldsymbol{u} : [0, \infty) \to \mathbb{R}^m$ is the control, $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ is the Chebyshev differentiation matrix, $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, $\boldsymbol{R} \in \mathbb{R}^{m \times m}$ are diagonal positive definite matrices, and "$\circ$" denotes elementwise multiplication. The parameters $\nu, \beta > 0, \boldsymbol{\alpha} \in \mathbb{R}^n$, and $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ are defined in Nakamura-Zimmerer et al. (2021b), and we take $n = 64$ and $m = 2$.

We train the NNs using supervised learning, which consists of three steps: data generation, optimization, and evaluation against test data. For details we refer the reader to Nakamura-Zimmerer et al. (2021b, 2022b).

We generate data by solving the BVP (6) for randomly sampled initial conditions. To get models with varying approximation accuracy, we generate training data sets with different numbers of trajectories. To account for statistical variation, for each different data set size we conduct ten trials with different randomly generated training trajectories and NN weight initializations. We use an independent test data set containing 500 trajectories.

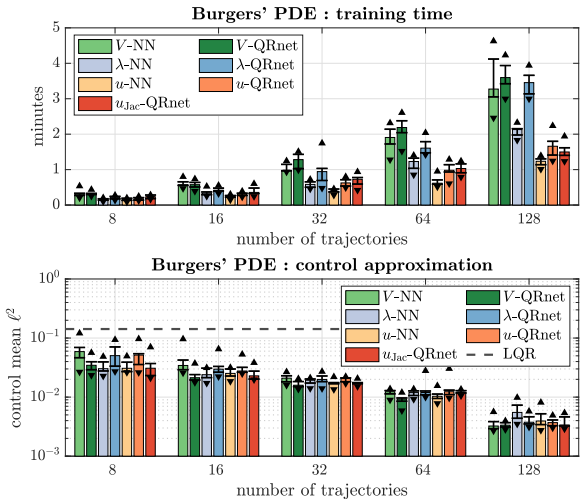


Fig. 2. Training time and test approximation error, depending on the amount of training data. Bar heights show the medians over ten trials, error bars show the 25th and 75th percentiles, and triangles are minimum and maximum values.

All NNs use five hidden layers with 32 neurons each. Figure 2 shows training times and test accuracies of the NNs. We see that, somewhat surprisingly, *u_{Jac}-QRnet* is just as fast to train as *u-NN* and *u-QRnet*[2]. We also find that *u_{Jac}-QRnet* has similar test accuracy statistics to the standard NNs, confirming that they can learn complicated nonlinear functions in practice.

Remark 3. Figures 3 to 5 show results for u-NN, *u-QRnet*, and *u_{Jac}-QRnet*. Results for other models are similar so we restrict the figures to these for clarity.

4.1 Local stability verification

As a first step we assess the local stability of each NN-controlled system. Let $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ be an equilibrium[3] of the closed loop system. The dynamics near $\bar{\boldsymbol{x}}$ can be approximated by $\dot{\boldsymbol{x}} \approx \boldsymbol{A}_{\text{cl}}(\boldsymbol{x} - \bar{\boldsymbol{x}})$, where

$$\boldsymbol{A}_{\text{cl}} := \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\right|_{\bar{\boldsymbol{x}}, \widehat{\boldsymbol{u}}(\bar{\boldsymbol{x}})} + \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right|_{\bar{\boldsymbol{x}}, \widehat{\boldsymbol{u}}(\bar{\boldsymbol{x}})} \left.\frac{\partial \widehat{\boldsymbol{u}}}{\partial \boldsymbol{x}}\right|_{\bar{\boldsymbol{x}}} \tag{13}$$

is the closed loop Jacobian. Thus, after synthesizing a feedback controller we can easily check for local stability by seeing if $\boldsymbol{A}_{\text{cl}}$ is Hurwitz.

Figure 3 shows the real part of the most positive eigenvalue of $\boldsymbol{A}_{\text{cl}}$ for each NN. We find that standard NNs must be trained to a high level of accuracy before they are even locally stabilizing, which necessitates a large data set and

[2] Some other novel architectures proposed by Nakamura-Zimmerer et al. (2022b), for example *λ_{Jac}-QRnet* which approximates the value gradient, do take significantly longer to train than standard NNs. However, even for these the training time is still very reasonable.

[3] In general $\bar{\boldsymbol{x}} \neq \boldsymbol{x}_f$. An equilibrium $\bar{\boldsymbol{x}}$ near $\boldsymbol{x}_f$, if it exists, can be obtained with a root-finding algorithm.

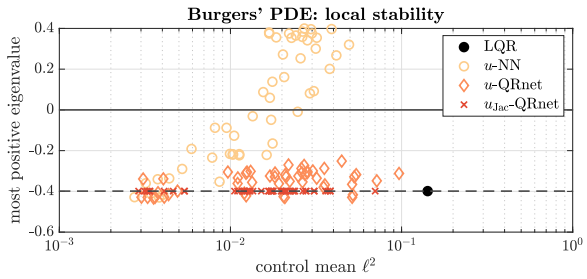


Fig. 3. Most positive real part of closed loop Jacobian eigenvalue. Each marker represents a single model.

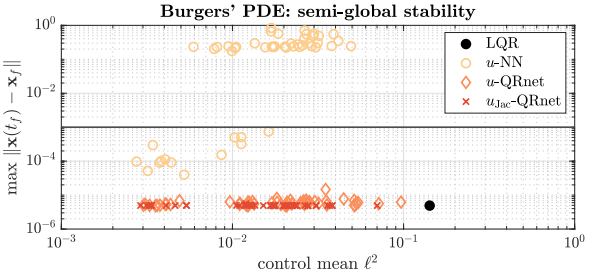


Fig. 4. Worst-case failure of final state over $N_{\text{MC}} = 100$ simulations. Each marker represents a single model.

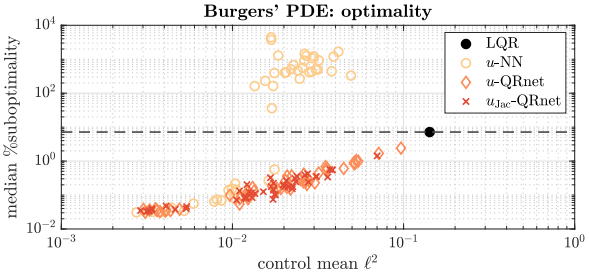


Fig. 5. Percent extra cost vs. BVP data over $N_{\text{MC}} = 100$ simulations. Each marker represents a single model.

long training time. On the other hand, *u-QRnet* and u_{Jac}-*QRnet* yield LAS even when trained on small data sets (recall Proposition 1 guarantees this for u_{Jac}-*QRnet*).

4.2 Monte Carlo nonlinear stability analysis

Here and in Section 4.3 we conduct Monte Carlo (MC) simulations to analyze nonlinear closed loop stability and performance. $N_{\text{MC}} = 100$ initial conditions $\boldsymbol{x}_0^{(i)}$ are randomly selected with $\left\|\boldsymbol{x}_0^{(i)}\right\| = 1.2 \approx \max_{\boldsymbol{x}^{(j)} \in \mathcal{D}_{\text{train}}} \left\|\boldsymbol{x}^{(j)}\right\|$, placing them at the edge of the training domain where the NNs may be less accurate.

We stop each simulation when the system reaches a steady state or exceeds a large final time. If the *worst case failure*, $\max_{\boldsymbol{x}_0^{(i)}} \left\|\boldsymbol{x}\left(t_f; \boldsymbol{x}_0^{(i)}\right)\right\|$, is sufficiently small then the nonlinear system is likely semi-globally stable. Conversely, if the controller fails to stabilize even one trajectory then we cannot rely on it.

Figure 4 shows the worst-case failures for each NN. Notably, only the most accurate *u*-NNs successfully stabilize the system. Furthermore, although stability on average improves with test accuracy, some highly accurate *u*-NNs fail to stabilize all trajectories. In contrast, *all u-QRnets* and u_{Jac}-*QRnet* stabilize all MC trajectories. These empirical results suggest that the proposed architectures make the control design more reliable, consistently yielding a stabilizing control law even with small data sets.

4.3 Monte Carlo optimality analysis

In this talk we are interested in both stability and optimality. For a given $\widehat{\boldsymbol{u}}(\cdot)$, optimality can be quantified by the accumulated cost compared to the optimal costs, $V\left(\boldsymbol{x}_0^{(i)}\right)$. Figure 5 shows the results of this analysis for the same set of MC simulations conducted in Section 4.2. We see a clear correlation between higher test accuracy and better performance. All the NN controllers (if stable) follow this trend, and moreover, their average performance is better than LQR alone. It follows that the proposed architectures improve stability without sacrificing optimality.

Ultimately, these results support the argument that while machine learning can serve as an effective computational tool, it is crucial to integrate control theory at each step of the model design, training, and evaluation process.

REFERENCES

Azmi, B., Kalise, D., and Kunisch, K. (2021). Optimal feedback law recovery by gradient-augmented sparse polynomial regression. *J. Mach. Learn. Res.*, 22(48), 1–32.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, 4(2), 251–257. doi:10.1016/0893-6080(91)90009-T.

Izzo, D. and Öztürk, E. (2021). Real-time guidance for low-thrust transfers using deep neural networks. *J. Guid. Control Dyna.*, 1–13. doi:10.2514/1.G005254.

Izzo, D., Tailor, D., and Vasileiou, T. (2021). On the stability analysis of deep neural network representations of an optimal state-feedback. *IEEE Trans. Aerosp. Electron. Syst.*, 57(1), 145–154. doi:10.1109/TAES.2020.3010670.

Kang, W. and Wilcox, L.C. (2017). Mitigating the curse of dimensionality: Sparse grid characteristics method for optimal feedback control and HJB equations. *Comput. Optim. Appl.*, 68(2), 289–315. doi:10.1007/s10589-017-9910-0.

Nakamura-Zimmerer, T., Gong, Q., and Kang, W. (2021a). Adaptive deep learning for high-dimensional Hamilton–Jacobi–Bellman equations. *SIAM J. Sci. Comput.*, 43(2), A1221–A1247. doi:10.1137/19M1288802.

Nakamura-Zimmerer, T., Gong, Q., and Kang, W. (2021b). QRnet: Optimal regulator design with LQR-augmented neural networks. *IEEE Control Syst. Lett.*, 5(4), 1303–1308. doi:10.1109/LCSYS.2020.3034415.

Nakamura-Zimmerer, T., Gong, Q., and Kang, W. (2022a). Neural network optimal feedback control with enhanced closed loop stability. In *American Control Conference (ACC, to appear)*, 2373–2378. doi:10.48550/arXiv.2109.07466.

Nakamura-Zimmerer, T., Gong, Q., and Kang, W. (2022b). Neural network optimal feedback control with guaranteed local stability.

Tailor, D. and Izzo, D. (2019). Learning the optimal state-feedback via supervised imitation learning. *Astrodynamics*, 3(4), 361–374. doi:10.1007/s42064-019-0054-0.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Extended Abstract : Viability and Invariance of Systems on Metric Spaces

Zeinab Badreddine and Hélène Frankowska

CNRS, IMJ-PRG, Sorbonne Université, 75252 Paris, France
(e-mail: zeinab.badreddine@imj-prg.fr; helene.frankowska@imj-prg.fr)

Abstract: In some applied models (as for instance of flocking or of the crowd control) it is more natural to deal with elements of a metric space (as for instance a family of subsets of a vector space endowed with the Hausdorff metric) rather than with vectors of a normed vector space. We consider a generalized control system on a metric space and investigate necessary and sufficient conditions for viability and invariance of proper subsets, describing state constraints. As examples of application we study controlled continuity equations on the metric space of probability measures, endowed with the Wasserstein distance, and controlled morphological systems on the space of nonempty compact subsets of the Euclidean space endowed with the Hausdorff metric. We also provide sufficient conditions for the existence and uniqueness of contingent solutions to the Hamilton-Jacobi-Bellman equation on a proper metric space.

Keywords: mutational control system; viability and invariance; optimal control; Wasserstein space; Hamilton-Jacobi inequalities.
MSC 2020: 34A06, 49L12, 49Q22.

Dynamical systems *under state constraints* are ubiquitous in the literature since a long while. Indeed, in many applied fields, as economics, finance, demography, medical sciences, aerospace, sustainable development, robotics, etc., the models do involve pointwise constraints on trajectories. Viability theory is an area of mathematics that studies evolutions of trajectories of dynamical systems under state constraints and many related questions. This theory is also a helpful tool for investigation of some classical questions arising in control. Indeed, various problems of control theory can be linked to viability and invariance properties of appropriately chosen sets, as for instance the optimal synthesis problem can be related to the *viability retroaction map* on the epigraph of the value function arising in optimal control [10], solutions of Hamilton-Jacobi-Bellman equations – to functions having viable/invariant epigraph/hypograph under extended control systems [11], optimal trajectories – to viable trajectories on the epigraph of the value function for an extended control system [11], stabilising controls – to the viability retroaction map on the epigraph of a Lyapunov function [12]. Viability theory is well investigated in the finite dimensional framework and Hilbert spaces, see for instance [4, 5, 8] and the bibliographies contained therein.

In the recent years, there is an increasing interest in control problems stated on metric spaces, cf. [2, 9, 13, 14]. The main goal of the present paper is to extend the viability and invariance theorems to the framework of proper subsets of a metric space on which a subset of transitions is fixed.

⋆ This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-18-1-0254.

As an application, we discuss the Hamilton-Jacobi inequalities on a proper metric space. We show that under some technical assumptions the value function of the mutational Mayer optimal control problem is the unique contingent solution satisfying a prescribed final time condition. Two examples of applications are provided as well.

1. PRELIMINARIES AND BASIC DEFINITIONS

Let (E,d) be a metric space (with the metric d) and denote by $B(x,r)$ the closed ball centered at $x\in E$ with radius $r\geq 0$. Recall that a subset $K\subset E$ is called proper if $K\cap B(x,r)$ is compact for any $(x,r)\in K\times\mathbb{R}_+$. The distance between two nonempty subsets K, M of E is defined by $\mathrm{dist}(K,M):=\inf_{k\in K,\,m\in M} d(k,m)$. Note that $\mathrm{dist}(K,M)$ measures the proximity of sets and it is not a distance function on subsets of E. Obviously, it is smaller than the Hausdorff distance.

We first recall some definitions and notations from [3, 15].

A map $\mathcal{V}:[0,1]\times E\to E$ is called transition on (E,d) if:

- $\forall\, x\in E,\ \mathcal{V}(0,x)=x$;
- $\forall\, x\in E,\ \forall\, t,\ h\in[0,1[\ \text{ with }\ t+h\leq 1,$ $\mathcal{V}(t+h,x)=\mathcal{V}(h,\ \mathcal{V}(t,x))$;
- $\alpha(\mathcal{V}):=\sup\limits_{x,y\in E;x\neq y}\ \limsup\limits_{h\to 0+}\dfrac{d(\mathcal{V}(h,x),\mathcal{V}(h,y))-d(x,y)}{h\,d(x,y)}$ $<+\infty$;
- $\beta(\mathcal{V}):=\sup\limits_{x\in E}\limsup\limits_{h\to 0+}\dfrac{d(x,\ \mathcal{V}(h,x))}{h}<+\infty$.

For any transitions $\mathcal{V}_1$, $\mathcal{V}_2$ on (E,d), define the pseudo distance

$$d_\Lambda(\mathcal{V}_1,\mathcal{V}_2):=\sup_{x\in E}\limsup_{h\to 0+}\frac{1}{h}\,d(\mathcal{V}_1(h,x),\ \mathcal{V}_2(h,x)).$$

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Let $\Theta(E)$ be a fixed nonempty set of transitions, $T > 0$ and $x(\cdot) : [0, T] \to E$. For $t \in [0, T[$, the set

$$\overset{\circ}{x}(t) := \{\mathcal{V} \in \Theta(E) \,|\, \lim_{h\to 0+} \frac{1}{h}\, d\Big(\mathcal{V}(h, x(t)), x(t+h)\Big) = 0\}$$

is called mutation of $x(\cdot)$ at time t (relative to $\Theta(E)$).

A mapping $x(\cdot) : [0, T] \to E$ is called primitive of $\mathcal{V} : [0, T] \to \Theta(E)$ if it is Lipschitz and $\overset{\circ}{x}(t) \ni \mathcal{V}(t)$ for a.e. $t \in [0, T]$. The reverse sign $\ni$ reflects the fact that $\overset{\circ}{x}(t)$ may be multivalued, while $\mathcal{V}(t)$ is single-valued.

For a nonempty subset K of E and $x \in K$, the set

$$\overset{\circ}{T}_K(x) := \{\mathcal{V} \in \Theta(E) \,|\, \liminf_{h\to 0+} \frac{1}{h}\mathrm{dist}(\mathcal{V}(h, x), K) = 0\}$$

is called the contingent transition set to K at x.

Let $T > 0$, $W : [0, T] \times E \to \mathbb{R} \cup \{+\infty\}$ and (t, x) be in the domain of W with $t < T$. Below, $\mathbf{0} : [0, 1] \times \mathbb{R} \to \mathbb{R}$ and $\mathbf{1} : [0, 1] \times \mathbb{R} \to \mathbb{R}$ denote the transitions on $\mathbb{R}$ defined by $\mathbf{0}(h, t) = t$, $\mathbf{1}(h, t) = t + h$, for all $h \in [0, 1]$, $t \in \mathbb{R}$. For any transition $\mathcal{V} \in \Theta(E)$, the contingent directional derivative of W at (t, x) in the direction $(\mathbf{1}, \mathcal{V})$ is defined by

$$\overset{\circ}{D}_{\uparrow} W(t, x)(\mathbf{1}, \mathcal{V}) = \lim_{\varepsilon\to 0+} \inf_{\substack{h\in]0,\varepsilon] \\ y\in B(\mathcal{V}(h,x),\varepsilon h)}} \frac{W(t + h, y) - W(t, x)}{h}$$

As in [6], it is not difficult to link directional derivatives to the contingent transition set to the epigraph of W at (t, x).

Let (U, d_U) be a complete separable metric space of control parameters and define the set of admissible controls by

$$\mathcal{U} := \{u(\cdot) : [0, \infty) \to U \,|\, u(\cdot) \text{ is Lebesgue measurable}\}.$$

2. WEAK INVARIANCE, VIABILITY AND INVARIANCE

Consider a map $f : E \times U \to \Theta(E)$. f is called continuous if for any $(x_0, u_0) \in E \times U$ and $\varepsilon > 0$, there exists some $\delta > 0$ such that for all $(x, u) \in E \times U$,

$$d(x, x_0) + d_U(u, u_0) < \delta \implies d_\Lambda(f(x, u), f(x_0, u_0)) < \varepsilon.$$

$f(\cdot, u)$ is said to be uniformly Lipschitz in u, if there is a constant $L > 0$ such that

$$d_\Lambda(f(x, u),\ f(y, u)) \le L\, d(x, y), \quad \forall\, x, y \in E,\ \forall\, u \in U.$$

For any $x_0 \in E$ consider the mutational control system

$$[S] \quad \overset{\circ}{x}(s) \ni f(x(s),\ u(s)) \ \text{ a.e.},\ x(0) = x_0,\ u(\cdot) \in \mathcal{U}.$$

A Lipschitz mapping $x(\cdot) : [0, T] \to E$ is said to be a solution to [S] on $[0, T]$ for some $T > 0$, if there exists a control $u(\cdot) \in \mathcal{U}$ such that $\overset{\circ}{x}(t) \ni f(x(t), u(t))$ a.e. in $[0, T]$ and $x(0) = x_0$. If $f(\cdot, u)$ is uniformly Lipschitz in u, then, by [15, Proposition 21, p. 41] and the Gronwall lemma, to every control $u(\cdot) \in \mathcal{U}$ corresponds at most one solution of [S]. Below we always assume that f is continuous and that

$$\sup_{x\in E, u\in U} \alpha(f(x, u)) < +\infty; \quad \sup_{x\in E, u\in U} \beta(f(x, u)) < +\infty.$$

We also assume that for each $x_0 \in E$ and $\bar{u} \in U$ there exists a solution to $\overset{\circ}{x} \ni f(x, \bar{u})$, $x(0) = x_0$ defined on $[0, 1]$. By [15, Theorem 20, p.40] it is always the case when E is proper. For more general metric spaces such existence depends on the choice of $\Theta(E)$ and f.

For any $x_0 \in E$ and $t \ge 0$, denote by $\mathcal{S}_t(x_0)$ the set of all solutions to the mutational control system $[S]$ defined on $[0, t]$ and by $R(t; x_0) := \{x(t) \in E \,|\, x(\cdot) \in \mathcal{S}_t(x_0)\}$, the reachable set from x_0 at time t.

Consider the controlled mutational equation

$$\overset{\circ}{x}(s) \ni f(x(s),\ u(s)) \text{ a.e. } s \ge 0,\ u(\cdot) \in \mathcal{U} \qquad (1)$$

and let $K \subset E$ be a proper nonempty set. K is called weakly invariant under (1) if for every $x_0 \in K$ we have $R(t; x_0) \cap K \neq \emptyset$ for any $t \ge 0$, see [17].

K is called viable under (1) if for any $x_0 \in K$, there exists a solution $x(\cdot)$ of (1) with $x(0) = x_0$ satisfying $x(t) \in K$ for all $t \ge 0$.

K is called invariant under (1) if every solution $x(\cdot)$ of (1) with $x(0) \in K$ satisfies $x(t) \in K$ for all $t \ge 0$.

Clearly any viable set is weakly invariant. To illustrate that weak invariance does not yield viability, consider the two dimensional control system $x' = u(t)$ in $\mathbb{R}^2$ with $U = \{0\} \times \{-1, 1\}$ and the set K equal to the unit sphere in $\mathbb{R}^2$. Then every trajectory of this system staring in K leaves it immediately. At the same time $x_0 \in R(t; x_0)$ for any $x_0 \in K$ and $t \ge 0$. Therefore K is weakly invariant.

Proposition 1. (Necessary condition for weak invariance). Assume that (U, d_U) is compact and $K \subset E$ is weakly invariant. If $x \in K$ is so that for a sequence $h_i \to 0+$

$$\sup_{y\in R(h_i;x)} \inf_{u\in U}\ d\Big(y, f(x, u)(h_i, x)\Big) = o(h_i), \qquad (2)$$

then there exists $u \in U$ satisfying $f(x, u) \in \overset{\circ}{T}_K(x)$. In particular, if for every $x \in K$ we can find $h_i \to 0+$ satisfying (2), then $f(x, U) \cap \overset{\circ}{T}_K(x) \neq \emptyset$ for every $x \in K$.

Theorem 2. (Sufficient condition for weak invariance). Assume that $f(\cdot, u)$ uniformly Lipschitz in u, that $R(t; x_0)$ compact for all $x_0 \in K$ and $t > 0$ and that for any $x \in K$, $f(x, U) \cap \overset{\circ}{T}_K(x) \neq \emptyset$. Then K is weakly invariant under (1).

Theorem 3. (Sufficient condition for viability). Under the assumptions of Theorem 2 suppose that for each $x_0 \in K$ and $t > 0$, the set $\mathcal{S}_t(x_0)$ is closed in the metric of uniform convergence. Then K is viable under (1).

Under the assumptions of Theorem 2, it can be shown, using the Ascoli-Arzelà Theorem 47.1 from [16], that every sequence $\{x_n(\cdot)\}_n \subset \mathcal{S}_T(x_0)$ has a subsequence converging uniformly to a Lipschitz function $x : [0, T] \to E$. Clearly, $x(t) \in R(t; x_0)$ for all $t \in [0, T]$. This does not yield however that $x(\cdot) \in \mathcal{S}_T(x_0)$. When E is a locally compact, complete metric space, using a more sophisticated construction as in [17, Theorem of Barbashin], it is even possible to get a Lipschitz mapping $x : [0, T] \to E$ satisfying $x(t_2) \in R(t_2; t_1, x(t_1))$ for all $0 \le t_1 < t_2 \le T$. Still this does not imply that $x(\cdot) \in \mathcal{S}_T(x_0)$. To illustrate that the assumptions in the above two theorems are not equivalent, consider a compact subset $U \subset \mathbb{R}^n$ and the control system $x' = u(t) \in U$, $x(0) = x_0$. It is well known that its reachable sets are compact, while, in general, $\mathcal{S}_T(x_0)$ is not closed in the metric of uniform convergence.

Theorem 4. (Mutational invariance). Suppose that $f(\cdot, u)$ is uniformly Lipschitz in u. Then K is invariant under (1) if and only if $f(x, U) \subset \overset{\circ}{T}_K(x)$ for each $x \in K$.

3. HAMILTON-JACOBI INEQUALITIES

In this section, (E,d) is a proper metric space. Given an extended lower semicontinuous cost function $g : E \to \mathbb{R} \cup \{+\infty\}$ and $x_0 \in E$, we associate to it the Mayer optimal control problem

$[P]$ minimize $g(x(1))$

over all the solutions of the mutational control system [S] defined on $[0,1]$.

Consider the mutational equation

$$\overset{\circ}{z}(\cdot) \ni f(z(\cdot), u(\cdot)), \quad u(\cdot) \in \mathcal{U} \tag{3}$$

and define the value function $V : [0,1] \times E \to \mathbb{R} \cup \{+\infty\}$ by : for all $t \in [0,1]$ and $x \in E$, $V(t,x) := \inf\{g(z(1)) \,|\, z(\cdot)$ is a solution to (3) on $[t,1]$, $z(t) = x\} \in \mathbb{R} \cup \{\pm\infty\}$.

Theorem 5. Assume that for any $x \in E$, the set $\mathcal{S}_1(x)$ is closed in the metric of uniform convergence. Then for any $t_0 \in [0,1]$ and $x \in E$ there exist a control $u(\cdot) \in \mathcal{U}$ and a solution z to (3) with $z(t_0) = x$ defined on $[t_0,1]$ and satisfying $V(t_0,x) = g(z(1))$.

Theorem 6. Assume that $f(\cdot,u)$ is uniformly Lipschitz in u and that for any $x \in E$, there exist $h_i \to 0+$ satisfying (2). Then, the value function V verifies the boundary condition $V(1,\cdot) = g(\cdot)$ and the following contingent inequalities:

- for any (t,x) in the domain of V with $t < 1$, $\sup_{u \in U} \overset{\circ}{D}_{\uparrow}(-V)(t,x)(\mathbf{1}, f(x,u)) \le 0$.
- for any (t,x) in the domain of V with $t < 1$, $\overset{\circ}{D}_{\uparrow} V(t,x)(\mathbf{1}, f(x,u)) \le 0$ for some $u \in U$.

A continuous map $w : [0,1] \times E \to \mathbb{R}$ is called a contingent solution to the mutational Hamilton-Jacobi equation (associated with $[P]$, $[S]$) if it satisfies the boundary condition $w(1,\cdot) = g(\cdot)$ and the above two contingent inequalities with V replaced by w.

Theorem 7. Assume that $f(\cdot,u)$ is uniformly Lipschitz in u. If g is continuous, then V is continuous. Furthermore, if g is locally Lipschitz, then V is locally Lipschitz.

Theorem 8. Let $g : E \to \mathbb{R}$ be continuous and $f(\cdot,u)$ be uniformly Lipschitz in u. Assume that for any $x \in E$, the set $\mathcal{S}_1(x)$ is closed in the metric of uniform convergence and there exist $h_i \to 0+$ satisfying (2). Then V is the unique continuous contingent solution to the mutational Hamilton-Jacobi equation.

4. EXAMPLES

In this section we discuss two examples where the general results of previous sections do apply. We endow the space $Lip(\mathbb{R}^N, \mathbb{R}^N)$ of all bounded Lipschitz continuous functions $F : \mathbb{R}^N \to \mathbb{R}^N$ with the topology of local uniform convergence. For any $F \in Lip(\mathbb{R}^N, \mathbb{R}^N)$, denote by $Lip\, F$ the smallest Lipschitz constant of F and set $||F||_\infty := \sup_{x \in \mathbb{R}^N} |F(x)|$. For $F : \mathbb{R}^N \to \mathbb{R}^N$ and $K \subset \mathbb{R}^N$ define $(Id + F)(K) := \{x + F(x) \mid x \in K\}$.

4.1 Morphological control problem

Consider the metric space $\mathcal{K}(\mathbb{R}^N)$ of nonempty compact subsets of $\mathbb{R}^N$ supplied with the Pompeiu-Hausdorff distance: for all $K_1, K_2 \in \mathcal{K}(\mathbb{R}^N)$

$$d_H(K_1,K_2) := \max\left\{ \max_{x \in K_1} \operatorname{dist}(x,K_2),\ \max_{x \in K_2} \operatorname{dist}(x,K_1) \right\}.$$

Then $(\mathcal{K}(\mathbb{R}^N),\ d_H)$ is a proper metric space, see for instance [15, Proposition 47, p.57].

For any $F : [0,\infty) \to Lip(\mathbb{R}^N, \mathbb{R}^N)$, $t \ge 0$ and $K_0 \in \mathcal{K}(\mathbb{R}^N)$, the set $\mathcal{V}_{F(\cdot)}(t,K_0) := \{x(t) \,|\, x(\cdot) \in W^{1,1}\big([0,t],\mathbb{R}^N\big),\ x'(s) = F(s)(x(s))$ a.e., $x(0) \in K_0\}$ is called the reachable set at time t of the system governed by $F(\cdot)(\cdot)$ from the initial condition K_0. The subset of transitions is given by

$$\Theta(\mathcal{K}(\mathbb{R}^N)) = \{\mathcal{V}_F \,|\, F \in Lip(\mathbb{R}^N, \mathbb{R}^N)\}.$$

Let $f : \mathcal{K}(\mathbb{R}^N) \times U \to Lip(\mathbb{R}^N, \mathbb{R}^N)$ be continuous with

$$\sup_{u \in U,\, K \in \mathcal{K}(\mathbb{R}^N)} \Big(Lip\ f(K,u) + ||f(K,u)||_\infty \Big) < +\infty\ . \tag{4}$$

Consider the morphological control system

$$[M] \qquad \overset{\circ}{K}(\cdot) \ni \mathcal{V}_{f(K(\cdot),u(\cdot))},\ u(\cdot) \in \mathcal{U}.$$

A map $K(\cdot) : [0,T] \to \mathcal{K}(\mathbb{R}^N)$, where $T > 0$, is called a solution to [M] if $K(\cdot)$ is Lipschitz continuous with respect to d_H and for some $u(\cdot) \in \mathcal{U}$ and for a.e. $t \in [0,T]$

$$\lim_{h \to 0+} \frac{1}{h} \cdot d_H\Big(\mathcal{V}_{f(K(t),u(t))}(h, K(t)), K(t+h) \Big) = 0.$$

Setting $F(t) := f(K(t),u(t))$, it follows that $K(\cdot)$ is the mutational primitive of $\mathcal{V}_{F(\cdot)}$ on $[0,T]$. The results from [15, pp. 388, 113, 74 and 24] imply

Proposition 9. For any $K_0 \in \mathcal{K}(\mathbb{R}^N)$, $T > 0$ and $u(\cdot) \in \mathcal{U}$, there exists a solution $K(\cdot)$ to $[M]$ on $[0,T]$ with $K(0) = K_0$ and for every time $t \in [0,T]$, $K(t)$ coincides with the reachable set of the differential equation

$$x'(\tau) = f\big(K(\tau), u(\tau)\big)\big(x(\tau)\big), \quad x(0) \in K_0.$$

Moreover, if $f(\cdot,u)$ is uniformly Lipschitz in u w.r.t. the metric on $Lip(\mathbb{R}^N, \mathbb{R}^N)$ generated by $\|\cdot\|_\infty$, then the solution to [M] with the initial condition K_0 is unique.

We shall need the following assumptions:

$$(\mathbf{H1}) \quad \begin{cases} (4) \text{ holds and } (U, d_U) \text{ is a compact metric space;} \\ f(\cdot,u) \text{ is uniformly Lipschitz in } u \text{ w.r.t. } \|\cdot\|_\infty; \\ f(K,U) \text{ is convex for every } K \in \mathcal{K}(\mathbb{R}^N). \end{cases}$$

Theorem 10. Assume **(H1)** and consider a closed nonempty subset $\Omega \subset \mathcal{K}(\mathbb{R}^N)$. Then Ω is viable under [M] if and only if for each $K \in \Omega$, there exists some $u \in U$ satisfying

$$\liminf_{h \to 0+} \frac{1}{h} \operatorname{dist}((Id + hf(K,u))(K), \Omega) = 0.$$

Furthermore, Ω is invariant under [M] if and only if the above equality holds true for each $K \in \Omega$ and any $u \in U$.

4.2 Control system in a Wasserstein space

Denote by $\mathcal{P}(\mathbb{R}^N)$ the family of all Borel probability measures on $\mathbb{R}^N$ endowed with the narrow topology.
For any $\emptyset \ne K \subseteq \mathbb{R}^N$, denote by $\mathcal{P}(K) \subseteq \mathcal{P}(\mathbb{R}^N)$ the set of all Borel probability measures with the support contained in K and by $\mathcal{P}_c(\mathbb{R}^N)$ the subset of all Borel probability measures with a compact support.

We first recall some notions in $\mathcal{P}(\mathbb{R}^N)$, see for instance [1].

For any $\mu \in \mathcal{P}(\mathbb{R}^N)$ and a Borel map $T : \mathbb{R}^N \to \mathbb{R}^N$, let $T_{\#}\mu \in \mathcal{P}(\mathbb{R}^N)$ denote the pushforward of μ through T:

$$T_{\#}\mu(B) := \mu(T^{-1}(B)) \text{ for any Borel set } B \subset \mathbb{R}^N.$$

Denote $\mathcal{P}_2(\mathbb{R}^N) := \{\mu \in \mathcal{P}(\mathbb{R}^N) \,|\, \int_{\mathbb{R}^N} |x|^2 \, \mathrm{d}\mu(x) < +\infty\}$ and by W_2 the Wasserstein distance on $\mathcal{P}_2(\mathbb{R}^N)$, see [1]. The space $(\mathcal{P}_2(\mathbb{R}^N), W_2)$ is called the Wasserstein space of order 2. It is well known that $(\mathcal{P}_2(\mathbb{R}^N), W_2)$ is complete and separable. Furthermore, for any nonempty compact $K \subset \mathbb{R}^N$, the set $\mathcal{P}(K)$ is compact in $(\mathcal{P}_2(\mathbb{R}^N), W_2)$.

Consider a continuous map $f : \mathcal{P}_2(\mathbb{R}^N) \times U \to Lip(\mathbb{R}^N, \mathbb{R}^N)$ such that $f(\cdot, u)$ is uniformly Lipschitz in u w.r.t. the metric on $Lip(\mathbb{R}^N, \mathbb{R}^N)$ generated by $\|\cdot\|_\infty$ and the controlled continuity equation

$$[C] \qquad \partial_t \mu(t) + \nabla(f(\mu(t), u(t)) \cdot \mu(t)) = 0, \quad u(\cdot) \in \mathcal{U}.$$

Given $T > 0$, an absolutely continuous map $\mu(\cdot) : [0, T] \to \mathcal{P}_2(\mathbb{R}^N)$ is a solution to [C] on $[0, T]$ if for some $u(\cdot) \in \mathcal{U}$ it solves

$$\partial_t \mu(t) + \nabla(f(\mu(t), u(t)) \cdot \mu(t)) = 0 \tag{5}$$

on $[0, T]$ in the sense of distributions.

The existence and the representation of solutions of the non-local continuity equation (5) for every initial condition in $\mathcal{P}_c(\mathbb{R}^N)$ were investigated in [7]. We shall assume:

$$(\mathbf{H2}) \quad \begin{cases} A_2 := \sup_{u \in U,\, \mu \in \mathcal{P}_2(\mathbb{R}^N)} Lip\, f(\mu, u) < \infty; \\ \rho_2 := \sup_{u \in U,\, \mu \in \mathcal{P}_2(\mathbb{R}^N)} ||f(\mu, u)||_\infty < +\infty; \\ (U, d_U) \text{ is a compact metric space;} \\ f(\mu, U) \text{ is convex } \forall\, \mu \in \mathcal{P}_2(\mathbb{R}^N). \end{cases}$$

Under assumptions (H2), $\mathcal{P}_c(\mathbb{R}^N)$ is invariant by solutions of [C], see [7].

Let $g \in Lip(\mathbb{R}^N, \mathbb{R}^N)$ and for any $\mu_0 \in \mathcal{P}_c(\mathbb{R}^N)$ and $h \in [0, 1]$, define $\mathcal{V}_g(h, \mu_0) := \mu(h)$, where $\mu(\cdot) : [0, 1] \to \mathcal{P}_2(\mathbb{R}^N)$ is the distributional solution to the continuity equation

$$\partial_t \mu(t) + \nabla(g \cdot \mu(t)) = 0, \quad \mu(0) = \mu_0. \tag{6}$$

Then $\mathcal{V}_g : [0, 1] \times \mathcal{P}_c(\mathbb{R}^N) \to \mathcal{P}_c(\mathbb{R}^N)$ is a transition on $(\mathcal{P}_c(\mathbb{R}^N), W_2)$. For any $t > 0$ and $\mu_0 \in \mathcal{P}_c(\mathbb{R}^N)$, consider the sets $\mathcal{S}_t^C(\mu_0)$ of solutions to [C] on $[0, t]$ with $\mu(0) = \mu_0\}$ and

$$R^C(t, \mu_0) := \{\mu(t) \,|\, \mu \in \mathcal{S}_t^C(\mu_0)\}.$$

It follows from [7] that assumption **(H2)** implies that for any $T > 0$ and $\mu_0 \in \mathcal{P}_c(\mathbb{R}^N)$ the set $\mathcal{S}_T^C(\mu_0)$ is compact in the metric of uniform convergence.

Proposition 11. Assume **(H2)**. Then, there exists $k > 0$ such that for any $\mu_0 \in \mathcal{P}_c(\mathbb{R}^N)$, any $h_n \to 0+$ and any $\mu(h_n) \in R^C(h_n, \mu_0)$ we can find $u_n \in U$ satisfying

$$W_2\Big(\mu(h_n), \mathcal{V}_{f(\mu_0, u_n)}(h_n, \mu_0)\Big) \le k h_n^2$$

and $W_2\Big((Id + f(\mu_0, u_n))_{\#}\mu_0, \mathcal{V}_{f(\mu_0, u_n)}(h_n, \mu_0)\Big) \le A_2 \rho_2 h_n^2$.

The above results together with those from Section 3 allow to deduce the following viability and invariance theorem.

Theorem 12. Assume **(H2)** and let $\Omega \subset \mathcal{P}_c(\mathbb{R}^N)$ be nonempty and proper. Then Ω is viable under [C] if and only if for each $\mu \in \Omega$, there exists some $u \in U$ satisfying

$$\liminf_{h \to 0+} \frac{1}{h} \operatorname{dist}((Id + h f(\mu, u))_{\#}\mu, \Omega) = 0.$$

Furthermore, Ω is invariant under [C] if and only if the above equality holds true for each $\mu \in \Omega$ and any $u \in U$.

REFERENCES

[1] L. Ambrosio, N. Gigli, G. Savaré, Gradient Flows in Metric Spaces and in the Space of Probability Measures, Birkhaüser, 2000.

[2] L. Ambrosio, J. Feng, *On a class of first order Hamilton-Jacobi equations in metric spaces*, J. Differential Equations, 256, 2194-2245, 2014.

[3] J.-P. Aubin, Mutational and Morphological Analysis: Tools for Shape Regulation and Morphogenesis, Birkhaüser, 1999.

[4] J.-P. Aubin, Viability Theory, Birkhaüser, 1991.

[5] J.-P. Aubin, A.-M. Bayen, P. Saint-Pierre, Viability Theory. New Directions, (2nd edition), Springer, 2011.

[6] J.-P. Aubin, H. Frankowska, Set-Valued Analysis, Birkhäuser, Boston, Basel, Berlin, 1990 (Modern Birkhäuser Classics, reprint 2008).

[7] B. Bonnet, H. Frankowska, *Differential inclusions in Wasserstein spaces: the Cauchy-Lipschitz framework*, Journal of Differential Equations, 271, 594-637, 2021.

[8] P. Cannarsa, G. Da Prato, H. Frankowska, *Domain invariance for local solutions of semilinear evolution equations in Hilbert spaces*, J. London Math. Soc., 102, 287–318, 2020.

[9] G. Cavagnari, A. Marigonda, B. Piccoli, *Generalized dynamic programming principle and sparse mean-field control problems*, Journal of Mathematical Analysis and Applications, 481, 2020.

[10] H. Frankowska, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, Applied Mathematics and Optimization, 19, 291-311, 1989.

[11] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equation*, SIAM J. on Control and Optimization, 31, 257-272, 1993.

[12] H. Frankowska, S. Plaskacz, *A measurable upper semicontinuous viability theorem for tubes*, J. of Nonlinear Analysis, TMA, 26, 565-582, 1996.

[13] W. Gangbo, A. Tudorascu, *On differentiability in the Wasserstein space and well-posedness for Hamilton-Jacobi equations*, Journal de Mathématiques Pures et Appliquées, 125, 119–174, 2019.

[14] N. Gozlan, C. Roberto, P.-M. Samson, *Hamilton-Jacobi equations on metric spaces and transport entropy inequalities*, Rev. Mat. Iberoam., 30, 133-163, 2014.

[15] T. Lorenz, Mutational Analysis: A Joint Framework for Cauchy Problems in and Beyond Vector Spaces, Springer-Verlag, 2010.

[16] J. R. Munkres, Topology, second Edition, Pearson Education Limited, 2014.

[17] E. Roxin, *Stability in general control systems*, Journal of Differential Equations, 1, 115-150, 1965.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Compositional features and neural network complexity in deep learning ⋆

Wei Kang * **Qi Gong** **

* *Naval Postgraduate School, Monterey, CA and University of California, Santa Cruz, CA, USA (e-mail: wkang@nps.edu).*
** *University of California, Santa Cruz, CA, USA (e-mail: qgong@ucsc.edu)*

Abstract: In this study, we explore the relationship between the complexity of neural networks and the internal compositional structure of the function to be approximated. The results shed light on the reason why using neural network approximation helps to avoid the curse of dimensionality.

Keywords: Non-linear control systems, deep learning, optimal control, power systems

1. INTRODUCTION

In this study, we explore the relationship between the complexity of neural networks and the internal compositional structure of the function to be approximated. The results shed light on the reason why using neural network approximation helps to avoid the curse of dimensionality (COD). In Section 2, we discuss the challenge of COD in feedback control. In Section 3, we introduce four compositional features that determine the complexity and error upper bound of neural network approximation for dynamical and control systems. In Section 4, several examples are given to illustrate the widely observed phenomenon in science and engineering that complicated functions are formed by the composition of simple ones.

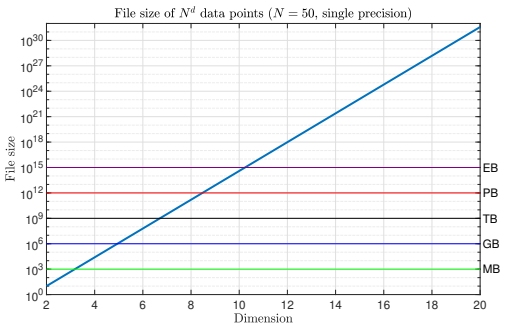


Fig. 1. The size of data file that grows exponentially with the space dimension. In each dimension, $N = 50$. The total number of data points is N^d.

2. THE CURSE OF DIMENSIONALITY

The COD is a bottleneck in many applications of dynamical systems and nonlinear control. It is a phenomenon in which the complexity of an approximate solution grows fast, such as exponentially, with the state space dimension. For instance, consider a feedback control law

$$u = u(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^d,\ u \in \mathbb{R}. \tag{1}$$

for a system of dimension d. If d is large and if an analytic representation cannot be found for $u(\mathbf{x})$, a numerical approximation has to be applied to store the control law in a digital format. If $u(\mathbf{x})$ is approximated using interpolation based on its value at grid points, the size of the dataset increases exponentially. Specifically, suppose we use N grid points in each dimension. Then the total number of grid points in a d dimensional domain is N^d. The value of $u(\mathbf{x})$ over the grid forms a huge dataset even for moderate dimensions such as $d = 7$ or 8. Figure 1 shows an example in which $N = 50$. If $d = 7$, the memory needed to store the value of $u(\mathbf{x})$ over the grid using single precision is about 1TB. This number is 1PB for $d = 9$ and 1EB for $d = 10$. Due to the limitations on processor's primary storage, bus speed and computation speed, the interpolation of datasets that have such large sizes is practically intractable, not to mention that the computation has to be carried out in real-time for feedback control.

3. COMPOSITIONAL FEATURES AND THE COMPLEXITY OF NEURAL NETWORKS

For the last few years, a new trend of overcoming the COD in nonlinear dynamics and control using deep learning has been developed rapidly. Many examples of successfully applying deep learning to high dimensional differential equations and optimal control were published in which the dimensions range from six to several hundred, well beyond what conventional computational methods can deal with (Han et al. (2018); Izzo et al. (2019); Nakamura-Zimmerer et al. (2021); Kang et al. (2021a); Sirignano and Spiliopoulos (2018); Raissi et al. (2019); B. Azmi (2020)). These empirical successes of deep learning in overcoming the COD inspire us to study the underlying reason why neural networks are capable of solving so many high dimensional problems. The philosophy in our study is based on a widely

⋆ This work was supported in part by U.S. Naval Research Laboratory - Monterey, CA and National Science Foundation

observed fact in science and engineering: complicated functions are formed by the composition of relatively simple functions. In Kang and Gong (2022), a set of key features of compositional functions is defined. It can be mathematically proved that these features determine the upper bounds of neural network complexity and approximation error. These upper bounds do not suffer from the COD.

To exemplify the compositional structure of nonlinear systems, consider the swing equations of a power system with N_g generators

$$\begin{aligned}\frac{d\omega_i}{dt} &= \frac{\omega_0}{2H_i}\left(P_m - D\frac{\omega_i-\omega_0}{\omega_0} - E_i^2 G_{ii} \cdots \right.\\ &\left. - \sum_{j=1,j\neq i}^{10} E_i E_j [B_{ij}\sin(\delta_i-\delta_j) + G_{ij}\cos(\delta_i-\delta_j)]\right)\\ \frac{d\delta_i}{dt} &= \omega_i - \omega_0,\end{aligned} \tag{2}$$

where $i = 1, ..., N_g$. For the ith generator, the two state variables are δ_i, the rotor angle in radian, and ω_i, the rotor speed in radian per second. Other parameters include H_i (the inertial constant of the generator), $\omega_0 = 2\pi \times f_0$ (the synchronous angular frequency in radian per second for an ac power system with frequency f_0), D (the damping coefficient), P_m (the mechanical power input from the turbine), E_i (the electromotive force or internal voltage of the generator). In addition, $G_{ij} + jB_{ij}$, the mutual admittance between E_i and E_j, is the i^{th} row j^{th} column element of the admittance matrix among all electromotive forces, and G_{ii} is the conductance representing the local load seen from E_i. Details about the model and its parameters refer to Athay et al. (1979).

A power system may have tens or hundreds of generators. This complicated system model, however, is a composition of functions that have low input dimensions. The compositional structure can be represented using a layered directed acyclic graph (DAG). For example, Figure 2 is a DAG of the function in (2). Each colored node in the DAG represents a nonlinear function. They are all sine and cosine functions with a single input. Although some linear nodes (white color) have high input dimensions, such as the node in the output layer, it is proved in Kang and Gong (2022) that linear nodes do not increase the complexity, or the number of neurons, of the neural network. The

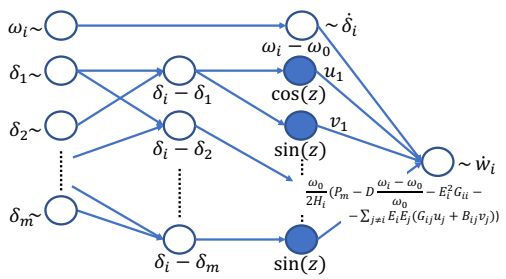


Fig. 2. The layered DAG of the function in (2) as a compositional function.

question we would like to answer is: how does a layered DAG help in the effort of using deep learning to find the trajectory or the optimal control of a system? Our study shows that some compositional features are critical to the required complexity of neural networks used in deep learning. These features are briefly introduced as follows.

- $|\mathcal{V}|$ (complexity feature): The total number of nonlinear nodes in the layered DAG, where $\mathcal{V}$ is the set of nonlinear nodes of the compositional function.
- r_{max} (dimension feature): The largest ratio, d/m, for all nodes in $\mathcal{V}$, where d is the input dimension of the node and m is the smoothness of the node.
- Λ (volume feature): Each nonlinear node, denote the function by f, has a domain. Assume that the domain is a square of edge length R. The volume feature is defined to be the largest value in
$$\{\max\{R,1\}||f||;\ f \in \mathcal{V}\}, \tag{3}$$
where $||\cdot||$ is the Sobolev norm
$$||f|| = ||f||_{L^\infty} + \sum \left\| \frac{\partial f}{\partial x_i} \right\|_{L^\infty}. \tag{4}$$
- $L_{\max}$ (Lipschitz constant feature): The largest Lipschitz constants associated with nonlinear nodes. Note that this Lipschitz constant is defined based on the layered structure of the DAG. For more details, the readers are referred to Kang and Gong (2022).

Consider a general dynamical system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^d \tag{5}$$

in which $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d$ has a layered compositional structure. Let $\boldsymbol{\phi}(t, \mathbf{x})$ represents the solution of (5) in which $\mathbf{x}$ is the initiate state, $\boldsymbol{\phi}(0, \mathbf{x}) = \mathbf{x}$.

Theorem 1. (Kang and Gong (2022)) Suppose that all nodes in $\mathbf{f}$ are C^1. Let $D \subset \mathbb{R}^d$ be a closed set and $R > 0$ be a constant. Suppose $\boldsymbol{\phi}(t, \mathbf{x}) \in [-R, R]^d$ for $t \in [0, T]$ and $\mathbf{x} \in D$. Then, there always exists a deep feedforward neural network, denoted by $\boldsymbol{\phi}^{NN}(\mathbf{x})$, in which activation functions are C^∞. Furthermore, the network satisfies

$$\left\|\boldsymbol{\phi}^{NN}(\mathbf{x}) - \boldsymbol{\phi}(T, \mathbf{x})\right\|_2 < (C_1 L_{max} \Lambda\, |\mathcal{V}| + C_2) n^{-1/r_{max}} \tag{6}$$

where n is an integer that determines an upper bound of the total number of neurons in $\boldsymbol{\phi}^{NN}(\mathbf{x})$, i.e., the complexity of the neural network,

$$\#\text{ of neurons in } \boldsymbol{\phi}^{NN} \leq \left(n^{1/r_{max}} + 1\right) n\, |\mathcal{V}| \tag{7}$$

The constants, C_1 and C_2, in (6) are determined by $\|\mathbf{f}\|_2$, $\left\|\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right\|_2$, T and the input dimensions of the nodes in $\mathcal{V}$.

It is worth to note that the error upper bound (6) depends on Λ, L_{max} and $|\mathcal{V}|$ as a polynomial function, rather than an exponential function. The value of r_{max} depends on the input dimensions of individual notes of $\mathbf{f}$, not directly on the overall dimension, d. Therefore, if r_{max} is bounded and if Λ, L_{max} and $|\mathcal{V}|$ do not increase exponentially with d, the neural network approximation of $\boldsymbol{\phi}(T, \mathbf{x})$ is free from the COD. A similar result holds true for optimal control. Consider a control system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}),\ \mathbf{x} \in D \subset \mathbb{R}^d,\ \mathbf{u} \in \mathbb{R}^q,\ t \in [0, T] \tag{8}$$

A zero-order hold control, $U = [\,\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_{N_t}\,]$ in which $\mathbf{u}_k$ is the constant control for $t \in [t_{k-1}, t_k]$. The goal of an optimal control problem is to find U that minimizes the cost function

$$J(\mathbf{x}, U) = \Psi \circ \boldsymbol{\phi}(\Delta t; \mathbf{u}_{N_t}, \cdot) \cdot \boldsymbol{\phi}(\Delta t; \mathbf{u}_{N_t-1}, \cdot) \circ \cdots \circ \boldsymbol{\phi}(\Delta t; \mathbf{u}_1, \mathbf{x}) \tag{9}$$

where $\Psi : \mathbb{R}^d \to \mathbb{R}$ is a function.

Theorem 2. Suppose that $\mathbf{f}$ and Ψ are compositional functions in which all nodes are C^2. Let $D \subset \mathbb{R}^d$ be a bounded closed set. Assume that the Hession of $J(\mathbf{x}, U)$ with respect to U is positive definite. Let $U^*(\mathbf{x})$ represents the optimal feedback control. Then, for any $\epsilon > 0$, there exists a deep neural network, U^{*NN}, that approximates the optimal control. The estimation error is

$$\left\| U^{*NN}(\mathbf{x}) - U^*(\mathbf{x}) \right\|_2 \leq 3\epsilon, \ \mathbf{x} \in D \tag{10}$$

The complexity of U^{*NN} is bounded by

$$n \leq C\epsilon^{-(4r_{max}+1+4r_{max}/r^{\mathbf{f}}_{max})} \tag{11}$$

where $r^{\mathbf{f}}_{max}$ is the dimension feature of $\mathbf{f}$ and r_{max} is the largest dimension feature of $\mathbf{f}$ and Ψ, C is a polynomial of q and other compositional features of $\mathbf{f}$ and Ψ.

4. SOME EXAMPLES OF COMPOSITIONAL FEATURES

According to Theorems 1 and 2, if the compositional features of a family of systems do not increase exponentially with d, the approximation of a trajectory or an optimal control has a polynomial error upper bound; the complexity of the neural network increases at a polynomial rate. In the following, we use three examples to demonstrate that this kind of polynomial relationship is not unusual.

4.1 Compositional features of power systems

The first example is the power system model in (2). Its layered DAG is shown in Figure 2. Its compositional features are summarized as follows.

$$r_{max} = 1, \ \ \Lambda = 4\pi, \ \ |\mathcal{V}| = 2(N_g - 1)N_g,$$
$$L_{max} = \max_{1 \leq i,j \leq N_g, i \neq j} \left\{ \frac{\omega_0}{2H_i} E_i E_j G_{ij}, \frac{\omega_0}{2H_i} E_i E_j B_{ij} \right\}. \tag{12}$$

The dimension feature is $r_{max} = 1$ because all nonlinear nodes ($\sin(z)$ and $\cos(z)$) have a single input. Here we treat the nodes as functions in C^1 although they are also in C^∞. This simplifies the formula in the derivation. The value of Λ depends on the radius of the domain of nonlinear nodes and their Sobolev norm (4). For each nonlinear node, the domain of its input is bounded by 2π. Furthermore, $\sin(z)$, $\cos(z)$ and their derivatives are all bounded by 1. Then it is straightforward to derive the volume feature $\Lambda = 4\pi$. Following the definition in Kang and Gong (2022), the Lipschitz constant associated with a node is the Lipschitz constant of $\mathbf{f}$ (not the node) with respect to the node when the node is treated as a free variable. For example, the nonlinear nodes in Figure 2 are $\cos(z)$ and $\sin(z)$. If one of them, for instance the first $\cos(z)$ connecting to $\delta_i - \delta_1$, is treated as a variable, the Lipschitz constant of the function associated with this node equals

$$\frac{\omega_0}{2H_i} E_i E_1 G_{i1}. \tag{13}$$

This computation is for the first nonlinear node ($j = 1$) in the ith generator. The value of L_{max} in (19) is the largest one among the numbers computed similarly for all the N_g generators and all nonlinear nodes. For the ith generator, $1 \leq i \leq N_g$, there are $2(N_g - 1)$ nonlinear nodes. Therefore, the total number of nonlinear nodes is $|\mathcal{V}| = 2(N_g - 1)N_g$. We would like to emphasize that the compositional features in (12) are either constants or polynomial functions of N_g. As such, they do not increase exponentially with N_g. From Theorem 1, there exists a deep neural network approximation of the power system that avoids the COD. In fact, a similar conclusion can be extended to the Lyapunov function of the power system, which is proved in Kang et al. (2021b).

Theorem 3. Consider a power system (2) that has N_g generators. Let $\mathcal{R} \subset \mathbb{R}^{2N_g}$ be a bounded set. Then, there exists a solution, $V(\mathbf{x})$, to Zubov's equation (a special Lyapunov function that characterizes the domain of attraction) and a neural network, $V^{NN}(\mathbf{x})$, that has n^{NN} hyperbolic tangent neurons. They satisfy

$$\left| V^{NN}(\mathbf{x}) - V(\mathbf{x}) \right| < (C_1 N_g^2 + C_2) \frac{N_g}{\sqrt{n^{NN}}} \tag{14}$$

for $\mathbf{x} \in \mathcal{R}$, where C_1 and C_2 are constants independent of N_g.

4.2 Lorenz-96 model

Consider a system of ODEs

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{15}$$

in which $\mathbf{f} : [-R, R]^d \subset \mathbb{R}^d \to \mathbb{R}^d$ is the vector field that defines the Loranz-96 model in Lorenz (1996),

$$\mathbf{f} = \begin{bmatrix} x_0(x_2 - x_{-1}) - x_1 + F \\ x_1(x_3 - x_0) - x_2 + F \\ \vdots \\ x_{i-1}(x_{i+1} - x_{i-2}) - x_i + F \\ \vdots \\ x_{d-1}(x_{d+1} - x_{d-2}) - x_d + F \end{bmatrix}, \tag{16}$$

where $x_{-1} = x_{d-1}$, $x_0 = x_d$, $x_{d+1} = x_1$ and F is a constant. Let's treat $\mathbf{f}$ as a compositional function. An example of its layered DAG when $d = 4$ is shown in Figure 3. All nonlinear nodes in Figure 3 are located in the second

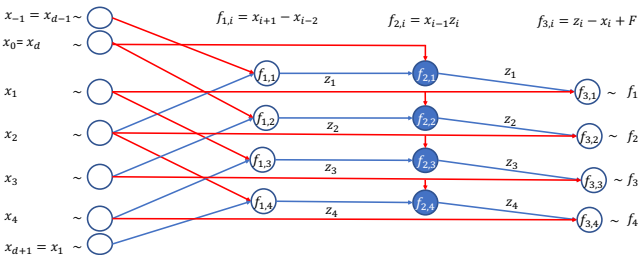


Fig. 3. DAG structure of the function (16) for $d = 4$. For a clear illustration, edges pointing to the first, the second, and the third layer are shown in red, blue and green respectively.

layer. They are defined by

$$f_{2,j}(x_{j-1}, z_j) = x_{j-1} z_j, \ \ j = 1, \cdots, d. \tag{17}$$

All nonlinear nodes have the dimension $d_{2,j} = 2$ and the domain $[-2R, 2R]^2$. Since $f_{2,j}$ is C^∞, we can set the smoothness to be any integer $m \geq 2$. As an example, we set $m = 1$. Then the dimension feature is $r_{max} = 2$. The Sobolev norm of $f_{2,j}$ is straightforward to compute, which

determines Λ, the volume feature. To compute the Lipchitz constant associated with the node, $f_{2,j}$, we construct the truncation of $\mathbf{f}$ along the second layer, which is given by

$$\bar{\mathbf{f}}(z_1, \cdots, z_{2d}) = \begin{bmatrix} z_1 - z_{d+1} + F \\ z_2 - z_{d+2} + F \\ \vdots \\ z_d - z_{2d} + F \end{bmatrix}, \tag{18}$$

where x_i, $1 \le i \le d$, are represented by the dummy inputs z_j, $j = d+1, \cdots, 2d$. The Lipschitz constant of $\bar{\mathbf{f}}$ with respect to z_j is $L_{2,j} = 1$, for $j = 1, \cdots, d$. The total number of nonlinear nodes in the system equals the dimension, d, because each equation in (16) has a single nonlinear node. To summarize, the compositional features of the Lorenz-96 model are

$$\begin{aligned} r_{max} &= 2,\ \Lambda = \max\{(2R), 1\}(2R + 4R^2), \\ L_{max} &= 1,\ |\mathcal{V}| = d. \end{aligned} \tag{19}$$

There is no exponential growth in the features. They are either constants or a linear function of d.

4.3 Burgers' Equation

Consider the following discretized Burgers's equation

$$\begin{aligned} \dot{u}_1 &= -u_1 \frac{u_2 - u_0}{2\Delta x} + \kappa \frac{u_2 + u_0 - 2u_1}{\Delta x^2} \\ \dot{u}_2 &= -u_2 \frac{u_3 - u_1}{2\Delta x} + \kappa \frac{u_3 + u_1 - 2u_2}{\Delta x^2} \\ &\vdots \\ \dot{u}_{N-1} &= -u_{N-1} \frac{u_N - u_{N-2}}{2\Delta x} + \kappa \frac{u_N + u_{N-2} - 2u_{N-1}}{\Delta x^2} \end{aligned} \tag{20}$$

The discretization is based on central different in which $\Delta x = L/N$ is the parameter that represents the step size of the state variable $x \in [0, L]$. The boundary condition is $u_0 = u_N = 0$. The dimension of the state space is N. The layered DAG of the function in (20) is shown in Figure 4. Due to the viscosity term in the equation, the solutions are stable. For initial conditions in a bounded set, we can assume that the state variables are bounded in $[-R, R]^N$ for some $R > 0$. The compositional features are summarized in (21). They are either constants or linear functions of N. None of them grows exponentially.

$$r^{\mathbf{f}}_{max} = 1, \Lambda^{\mathbf{f}} = \frac{1}{2}R^2 + R, L^{\mathbf{f}}_{max} = \frac{N}{L}, \left|\mathcal{V}^{\mathbf{f}}_G\right| = 2N. \tag{21}$$

5. CONCLUSION

The relationship revealed in this study between the complexity of neural networks and the compositional features in system models illustrates the reason why deep learning is an effective tool of overcoming the COD. The study raises more questions than answers. Many interesting problems are still widely open about the role of compositional structure in neural network design, as well as in the training and validation process.

ACKNOWLEDGEMENTS

This material is based upon activities supported by the National Science Foundation (under Interagency Agreement #2202668 and Award #2134235) and Naval Research Laboratory, Monterey, California. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation and Naval Research Laboratory.

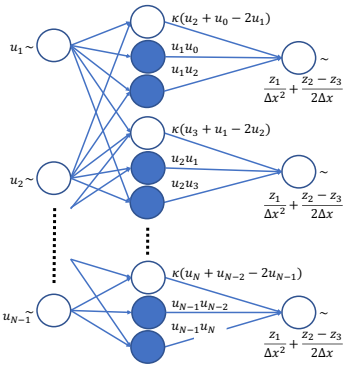


Fig. 4. DAG structure of the function (16). For a clear illustration, edges pointing to the first, the second, and the third layer are shown in red, blue and green respectively.

REFERENCES

Athay, T., Podmore, R., and Virmani, S. (1979). A practical method for the direct analysis of transient stability. *IEEE Transactions on Power Apparatus and Systems*, 98, 573–584.

B. Azmi, D. Kalise, K.K. (2020). ptimal feedback law recovery by gradient-augmented sparse polynomial regression. *arXiv:2007.09753*.

Han, J., Jentzen, A., and E., W. (2018). Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115, 8505–8510.

Izzo, D., Öztürk, E., and Märtens, M. (2019). Interplanetary transfers via deep representations of the optimal policy and/or of the value function. *arXiv:1904.08809*.

Kang, W. and Gong, Q. (2022). Neural network approximations of compositional functions with applications to dynamical systems. *SIAM Journal on Control and Optimization*, to applear.

Kang, W., Gong, Q., Nakamura-Zimmerer, T., and Fahroo, F. (2021a). Algorithms of data generation for deep learning and feedback design: A survey. *Physica D: Nonlinear Phenomena*, 425.

Kang, W., Sun, K., and Xu, L. (2021b). Data-driven computational methods for the domain of attraction and zubov's equation. *arXiv:2112.14415v1*.

Lorenz, E. (1996). Predictability – a problem partly solved. *ECMWF Seminar on Predictability*, I.

Nakamura-Zimmerer, T., Gong, Q., and Kang, W. (2021). Adaptive deep learning for high-dimensional Hamilton-Jacobi-Bellman equations. *SIAM J. Scientific Computing*, 43, 1221–1247.

Raissi, M., Perdikaris, P., and Karniadakis, G.E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.

Sirignano, J. and Spiliopoulos, K. (2018). DGM: A deep learning algorithm for solving partial differential equations. *arXiv:1708.07469*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Learning stability guarantees for data-driven constrained switching linear systems

Adrien Banse * **Zheming Wang** * **Raphaël Jungers** *

* *The ICTEAM Institute, UCLouvain, Louvain-la-Neuve,1348, Belgium (email: adrien.banse@student.uclouvain.be, zheming.wang@uclouvain.be, raphael.jungers@uclouvain.be).*

Abstract: We consider stability analysis of constrained switching linear systems in which the dynamics is unknown and whose switching signal is constrained by an automaton. We propose a data-driven Lyapunov framework for providing probabilistic stability guarantees based on data harvested from observations of the system. By generalizing previous results on arbitrary switching linear systems, we show that, by sampling a finite number of observations, we are able to construct an approximate Lyapunov function for the underlying system. Moreover, we show that the entropy of the language accepted by the automaton allows to bound the number of samples needed in order to reach some pre-specified accuracy.

Keywords: Stability analysis, Constrained switching linear systems, Data-driven optimization, Scenario approach

1. INTRODUCTION

In this paper we address the problem of finding probabilistic guarantees for the stability of *constrained switching linear systems* whose dynamics is unknown.

Switching systems. We consider discrete-time *switching linear systems* (*SLS*) defined by a set $\mathcal{A} = \{A_i, \}_{i\in\{1,\dots,m\}}$ of m matrices. Their dynamics is given by the following equation:

$$x_{t+1} = A_{\sigma(t)}x_t \tag{1}$$

for any $t \in \mathbb{N}$, where $x_t \in \mathbb{R}^n$ and $\sigma(t) \in \{1, \dots, m\}$ are respectively the *state* and the *mode* at time t. The sequence $(\sigma(0), \sigma(1), \dots) \subseteq \{1, \dots, m\}^{\mathbb{N}}$ is the *switching sequence*.

Switching linear systems are an important family of hybrid systems which often arise in Cyber-Physical systems (see Tabuada (2009)). Indeed, the interaction between continuous and discrete dynamics causes hybrid behaviors which makes the stability analysis challenging. In recent years, many model-based stability analysis techniques have been proposed (see Lin and Antsaklis (2009) and references therein, or Jungers (2009)).

A *constrained switching linear system* (*CSLS*) is a switching linear system with logical rules on its switching signal. We represent these rules by an *automaton*. The stability of CSLS has also been studied extensively (see e.g. Dai (2011), Philippe et al. (2016) and Xu and Acikmese (2020)). In particular, we are interested in asymptotic stability of CSLS, whose definition is given as follows. A CSLS whose dynamics is given by (1) is said to be *asymptotically stable* (or *stable*, for short) if for all $x_0 \in \mathbb{R}^n$,

$$\lim_{t\to\infty} x_t = 0. \tag{2}$$

Data-driven approach. In many practical applications, the engineer cannot rely on having a model, but rather has to analyze stability in a *data-driven* fashion. Most classical data-driven methods (see e.g. Karimi and Kammer (2017), Hjalmarsson et al. (1998) and Campi et al. (2003)) are limited to linear systems and based on classical identification and frequency-domain approaches. These methods may not be well suited for complex systems such as constrained switching linear systems.

In order to tackle hybrid behaviors in switching systems, novel data-driven stability analysis methods have been recently developed based on *scenario optimization* (see Kenanian et al. (2019), Berger et al. (2021) and Rubbens et al. (2021)). In this paper we seek to take one more step towards complexity. To do that, we develop a data-driven method for providing probabilistic guarantees on the stability of noise-free constrained switching linear systems.

Outline. The rest of this paper is organized in two parts. We introduce the problem that we tackle in Section 2. All concepts needed to this end are introduced in Section 2.1, and the problem is formulated in Section 2.2. In Section 3, we propose a lifting result allowing us to reduce the computation of the *constrained joint spectral radius* to the *joint spectral radius* of a certain set of matrices. Moreover, we state the main theorem of this paper, which extends data-driven results from Berger et al. (2021) to constrained switching linear systems. Finally, we investigate further the obtained generalization. We show that the notion of *entropy* can be used to characterize the number of samples needed to reach a specified guarantee on the stability. We will show that, under some assumptions, a smaller entropy allows for a better probabilistic guarantee.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2. PROBLEM SETTING

2.1 Preliminaries

In this subsection, we introduce the notions necessary to formally present the problem that we solve in this paper.

Joint spectral radius. For arbitrary SLS, given a set of matrices $\mathcal{A} = \{A_1, \dots, A_m\} \subseteq \mathbb{R}^{n\times n}$, the quantity

$$\rho(\mathcal{A}) = \lim_{t\to\infty} \max_{\sigma(\cdot)\in\{1,\dots,m\}} \|A_{\sigma(t-1)} \dots A_{\sigma(0)}\|^{1/t} \quad (3)$$

is known as the *joint spectral radius* (*JSR*) of a switching linear system defined on $\mathcal{A}$. The JSR of an switching system rules the stability of the latter:

Proposition 1. (Jungers (2009), Corollary 1.1.). Given a set of matrices $\mathcal{A}$, the switching linear system defined by $\mathcal{A}$ is asymptotically stable if and only if $\rho(\mathcal{A}) < 1$.

It is a well known fact that for any stable arbitrary switching linear system, there is a norm acting as a *common Lyapunov function* (see Jungers (2009), Proposition 1.4.). The following proposition gives a sufficient condition for stability, by restricting the search to *common quadratic Lyapunov functions* (*CQLF*).

Proposition 2. (Jungers (2009), Proposition 2.8). Consider a finite set of matrices $\mathcal{A}$. If there exists $\gamma \geq 0$ and a symmetric matrix $P \succ 0$ such that $A^T P A \preceq \gamma^2 P$ holds for any matrix $A \in \mathcal{A}$, then $\rho(\mathcal{A}) \leq \gamma$.

Constrained joint spectral radius. First, we give the definition of an *automaton*. An automaton is a strongly connected, directed and labelled graph $\mathbf{G}(V, E)$ with V the set of nodes and E the set of edges. Note that we drop the explicit writing of V and E when it is clear from the context. The edge $(v, w, \sigma) \in E$ between two nodes $v, w \in V$ carries the *label* $\sigma \in \{1, \dots, m\}$, which maps to a mode of the switching system. A sequence of labels $(\sigma(0), \sigma(1), \dots)$ is a *word* in the language *accepted* by the automaton $\mathbf{G}$ if there is a path in $\mathbf{G}$ carrying the sequence as the succession of the labels on its edges. A CSLS defined on the set of matrices $\mathbf{\Sigma}$ and constrained by the automaton $\mathbf{G}$ is noted $S(\mathbf{G}, \mathbf{\Sigma})$. We define the set of all possible products of matrices in $\mathbf{\Sigma}$ of length l given an automaton $\mathbf{G}$ as

$$\begin{aligned}\mathbf{\Pi}_l = \{&A_{\sigma(l-1)} A_{\sigma(l-2)} \dots A_{\sigma(0)} : \\ &(\sigma(0), \sigma(1), \dots, \sigma(l-1)) \text{ is a word of } \mathbf{G}\}.\end{aligned} \quad (4)$$

The constrained joint spectral radius (*CJSR*), which is a generalization of the JSR to CSLS, was first introduced in Dai (2011). Given a set of matrices $\mathbf{\Sigma}$ and an automaton $\mathbf{G}$, the CJSR of the constrained switching linear system $S(\mathbf{G}, \mathbf{\Sigma})$ is defined as

$$\rho(\mathbf{G}, \mathbf{\Sigma}) = \lim_{t\to\infty} \max \left\{ \|\mathbf{A}\|^{1/t} \,:\, \mathbf{A} \in \mathbf{\Pi}_t \right\}. \quad (5)$$

In the same way, the stability of a constrained switching linear system is characterized by its CJSR:

Proposition 3. (Dai (2011), Corollary 2.8.). Given a set of matrices $\mathbf{\Sigma}$ and an automaton $\mathbf{G}$, the constrained switching linear system $S(\mathbf{G}, \mathbf{\Sigma})$ is asymptotically stable if and only if $\rho(\mathbf{G}, \mathbf{\Sigma}) < 1$.

2.2 Problem formulation

We will now formally present the problem that we solve in this paper.

Model-based setting. Consider a given constrained switching linear system $S(\mathbf{G}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} \subseteq \mathbb{R}^{n\times n}$. Let $\Delta = \mathbb{S} \times \mathbf{\Pi}_l$ with $S \subseteq \mathbb{R}^n$ the unit sphere and $\mathbf{\Pi}_l$ the set of all admissible products of length l. We introduce the following optimization problem [1] (see Berger et al. (2021)):

$$\begin{aligned}\mathcal{P}(\Delta) : &\min_{\substack{P\in\mathbb{R}^{n\times n}\\ \gamma\geq 0}} (\gamma, ||P||_F^2) \\ &\text{s.t. } P \in \mathcal{X} := \left\{P : I \preceq P \preceq CI,\, P = P^T\right\}, \\ &\quad (\mathbf{A}x)^T P(\mathbf{A}x) \leq \gamma^{2l} x^T P x \quad \forall (x, \mathbf{A}) \in \Delta,\end{aligned} \quad (6)$$

for a large $C \in \mathbb{R}_{\geq 0}$, where $\|\cdot\|_F$ is the Frobenius norm. We denote $(\gamma^*(\Delta), P^*(\Delta))$ as the solution of optimization problem (6).

Following Proposition 2, Program (6) allows us to study stability in a model-based setting i.e., when Δ is known. Indeed if $\gamma < 1$, then the ellipsoidal norm $\|\cdot\|_{P^*(\Delta)}$ is a CQLF for the considered CSLS (Jungers, 2009). Observe that, in addition to the problem of Proposition 2, a tie-breaking rule is defined in Program (6). This tie-breaking rule allows for improving the probabilistic guarantees we obtain in Theorem 5 (see Kenanian et al. (2019) for details). A constraint $P \preceq CI$ is also added to ensure that the set of feasible P is compact, so that the existence of a solution is guaranteed [2].

Data-driven setting. In this work, we analyze the same problem in a data-driven framework: we assume that the system is not known (i.e., $\mathbf{A}$ is not known in Program (6)), but that we sample N trajectories of length l of a system $S(\mathbf{G}, \mathbf{\Sigma})$. The i-th trajectory is noted $(x_{i,0}, \dots, x_{i,l})$ for $i \in \{1, \dots, N\}$. The trajectories are assumed to be generated from initial states $x_{i,0}$ drawn randomly, uniformly and independently from $\mathbb{S}$, the unit sphere.

For each trajectory $i \in \{1, \dots, N\}$, the l matrices are generated from the automaton $\mathbf{G}(V, E)$ in the following way. An initial state u_0 is drawn randomly and uniformly from V. Then a random walk of length l is performed on $\mathbf{G}$, where, from $u_j \in V$, the next state u_{j+1} is drawn randomly, uniformly and independently from the set of its out-neighbours $\{u_{j+1} \in V : (u_j, u_{j+1}, \sigma_i(j) \in E\}$ where $\sigma_i(j)$ is the label corresponding to the edge linking u_j and u_{j+1}. The sequence of nodes $(u_0, \dots, u_j, u_{j+1}, \dots, u_l)$ form a switching sequence $\sigma_i(0), \dots, \sigma_i(l-1)$, which maps to the matrices $A_{\sigma_i(0)}, \dots, A_{\sigma_i(l-1)}$.

We define the set of N observations ω_N as

$$\omega_N = \{(x_{i,0}, \mathbf{A}_i), i = 1, \dots, N\} \quad (7)$$

where $\mathbf{A}_i = A_{\sigma_i(l-1)} \dots A_{\sigma_i(0)} \in \mathbf{\Pi}_l$. Note that the observations in ω_N are assumed to be noise-free.

[1] We note $\min(f(x), g(x))$ the multiobjective optimization problem where g(x) is used as a *tie-breaking rule*. That is, the objective is to minimize the function $f(x)$, and, in case there are several optimizers, the solution is the one which minimizes $g(x)$. Observe that the latter is unique because the problem is quasi-convex, and because $\|\cdot\|$ is a strongly convex function.

[2] For more details about these additions, see Berger et al. (2021).

We define $\mathbb{P} = \mathbb{P}_x \times \mathbb{P}_\sigma$ the probability measure on Δ with $\mathbb{P}_x$ the uniform distribution on $\mathbb{S}$ and $\mathbb{P}_\sigma$ the probability distribution describing the distribution of paths in $\mathbf{\Pi}_l$ as explained above. Note that $\mathbb{P}_\sigma$ is not necessarily a uniform measure.

Now, for a given set ω_N, let us define the *sampled optimization problem* $\mathcal{P}(\omega_N)$ associated to $\mathcal{P}$:

$$\begin{aligned}\mathcal{P}(\omega_N): \min_{\substack{P\in\mathbb{R}^{n\times n}\\ \gamma\geq 0}} \ & (\gamma, ||P||_F^2)\\ \text{s.t. } P\in\mathcal{X} &:= \left\{P : I \preceq P \preceq CI,\ P = P^T\right\},\\ (\mathbf{A}x)^T P(\mathbf{A}x) &\leq \gamma^{2l} x^T P x \quad \forall (x,\mathbf{A})\in\omega_N,\end{aligned} \tag{8}$$

We denote $(\gamma^*(\omega_N), P^*(\omega_N))$ as the solution of optimization problem (8), and $\text{Cost}(\omega_N)$ its optimal cost. The problem $\mathcal{P}(\omega_N)$ defined in Program (6) is the *data-driven* version of the optimization problem $\mathcal{P}(\Delta)$ defined in Program (8). The issue that we tackle in this paper is the inference of $\gamma^*(\Delta)$, the solution of optimization problem (6) from $(\gamma^*(\omega_N), P^*(\omega_N))$ with a certain user-defined level of confidence.

3. MAIN RESULTS

In this section, we present our main results. First, in Proposition 4 given an automaton $\mathbf{G}$ and a set of matrices $\mathbf{\Sigma}$, we show that the CJSR can be bounded by the classical JSR of the set of all admissible products of a given length $\mathbf{\Pi}_l$. Even though other reductions of the CJSR computation problems to a simpler JSR have already been proposed in the literature (see e.g. Dai (2011) and Philippe et al. (2016)), to the best of our knowledge, Proposition 4 is new, and will be useful for our purposes. Second, we use this result in order to derive a probabilistic guarantee allowing to relate the data-driven problem (8) to the model-based problem (6). This guarantee is given in Theorem 5.

Proposition 4. For all $l > 0$, given an automaton $\mathbf{G}$ and a set of matrices $\mathbf{\Sigma}$, the CJSR of $S(\mathbf{G}, \mathbf{\Sigma})$ and the JSR of the switching linear system defined by $\mathbf{\Pi}_l$ satisfy

$$\rho(\mathbf{G}, \mathbf{\Sigma}) \leq \rho(\mathbf{\Pi}_l)^{1/l}. \tag{9}$$

Moreover, the equality holds asymptotically i.e.,

$$\rho(\mathbf{G}, \mathbf{\Sigma}) = \lim_{l\to\infty} \rho(\mathbf{\Pi}_l)^{1/l}. \tag{10}$$

Proposition 4 allows us to reduce the problem of approximating the CJSR to the problem of approximating the JSR of another arbitrary switching linear system. Therefore we can generalize previous data-driven works on arbitrary systems. In particular, we draw our results on top of Berger et al. (2021) in order to obtain data-driven stability guarantees for constrained systems.

We remark that the data-driven problem (8) is a quasi-linear optimization problem, as defined in (Berger et al., 2021, Equation 1). Thus, a very similar analysis as in Berger et al. (2021), based on scenario-approach results Calafiore (2010), can be done. First, we recall the definition of a *Barabanov* matrix (see Berger et al. (2021), Defintion 7). A matrix $A \in \mathbb{R}^{n\times n}$ is said to be Barabanov if there exists a symmetric matrix $P \succ 0$ and $\gamma \geq 0$ such that $A^T P A = \gamma^2 P$.

Given Proposition 3, the following theorem generalizes Corollary 14 of Berger et al. (2021). It gives probabilistic guarantees for the stability of a constrained switching linear system. In the following theorem, $\Phi(\cdot, a, b)$ denotes the *regularized incomplete beta function* for the two parameters $a, b \in \mathbb{N}$ (see Kenanian et al. (2019), Definition 2).

Theorem 5. Consider an automaton $\mathbf{G}$, a set of matrices $\mathbf{\Sigma} \subseteq \mathbb{R}^{n\times n}$, samples $\omega_N \subset \Delta$ obtained as explained in Section 2.2, a fixed length $l > 0$ and $N \geq d := n(n+1)/2$. Suppose $\mathbf{\Pi}_l$ contains no Barabanov matrices. Consider problem $P(\omega_N)$ with solutions $\gamma^*(\omega_N)$ and $P^*(\omega_N)$. Then, for a given level of confidence $\beta \in (0,1)$,

$$\mathbb{P}\left(\left\{\omega_N \in \Delta^N : \rho(\mathbf{G}, \mathbf{\Sigma}) \leq \frac{\gamma^*(\omega_N)}{\sqrt[l]{\delta(\beta, \omega_N)}}\right\}\right) \geq \beta, \tag{11}$$

and the function $\delta(\beta, \omega_N)$ takes the form

$$\sqrt{1 - \Phi^{-1}(\varepsilon(\beta, N)\kappa(P^*(\omega_N))/p_{l,\min}, (n-1)/2, 1/2)}. \tag{12}$$

where $p_{l,\min}$ is the minimal probability of all matrices in $\mathbf{\Pi}_l$, $\kappa(P) = \sqrt{\det(P)/\lambda_{\min}(P)^n}$, and $\varepsilon(\beta, N)$ takes the closed form

$$\varepsilon(\beta, N) = 1 - \Phi(1-\beta, d+1, N-d). \tag{13}$$

Theorem 5 provides a general way of obtaining probabilistic stability guarantees. Indeed, for a given confidence level β, if one computes an upper bound (11) strictly less than 1, then following Proposition 3, stability holds with probability at least β.

We now show how one can use it in practice, by deriving a few corollaries. The following corollary holds if the distribution of drawing a product in $\mathbf{\Pi}_l$ is uniform.

Corollary 6. Suppose $\mathbb{P}_\sigma$ is a uniform measure. Then the function $\delta(\beta, \omega_N)$ in Theorem 5 can be written

$$\sqrt{1 - \Phi^{-1}(\varepsilon(\beta, N)|\mathbf{\Pi}_l|\kappa(P^*(\omega_N)), (n-1)/2, 1/2)}. \tag{14}$$

We now show that we can push further our analysis of the upper bound expressed in Corollary 6 by using the notion of *entropy* (Lind and Marcus, 1995, Definition 4.1.1). Let $|\mathcal{L}_{\mathbf{G},l}|$ be the language accepted by $\mathbf{G}$ restricted to length l. The entropy $h(\mathbf{G})$ of $\mathbf{G}$ is the growth rate of $|\mathcal{L}_{\mathbf{G},l}|$ i.e.,

$$h(\mathbf{G}) = \lim_{l\to\infty} \frac{\log_2 |\mathcal{L}_{\mathbf{G},l}|}{l}. \tag{15}$$

Since $|\mathbf{\Pi}_l| \leq |\mathcal{L}_{\mathbf{G},l}|$, the definition of the entropy gives the following corollary.

Corollary 7. For $l \to \infty$, the function $\delta(\beta, \omega_N)$ in Corollary 6 can be written

$$\lim_{l\to\infty} \sqrt{1 - \Phi^{-1}(\varepsilon(\beta, N)2^{lh(\mathbf{G})}\kappa(P^*(\omega_N)), (n-1)/2, 1/2)}. \tag{16}$$

Corollary 7 provides an asymptotic estimate of the probabilistic upper bound in Theorem 5, as a function of the entropy of the automaton $\mathbf{G}$. One can see that an automaton with small entropy allows for a better estimate of the CJSR, for a fixed number of samples. This is illustrated in Figure 1.

Now we show that we can also derive a practical bound for any finite $l > 0$, unlike Corollary 7 which holds

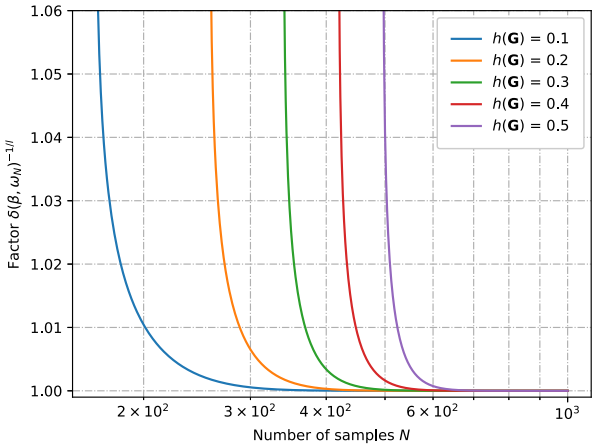


Fig. 1. Shape of the factor $1/\sqrt[l]{\delta(\beta, \omega_N)}$ in Theorem 5 with respect to the entropy, for a confidence level $\beta = 95\%$, a large l (here $l = 50$) and $n = 2$. One can see that this factor converges to 1 as N increases, and that a smaller entropy allows to converge faster.

asymptotically. For this we use classical results from graph theory.

Proposition 8. Let A be the adjacency matrix of some automaton $\mathbf{G}(V, E)$. Let $\lambda_1 \leq \cdots \leq \lambda_{|V|}$ be the eigenvalues of A. Assume A is diagonalizable. Then for any $l \geq 0$, $|\mathbf{\Pi}_l| \leq |V|\lambda_n^l$.

Proposition 8 directly gives the following corollary.

Corollary 9. Let A be the adjacency matrix of some automaton $\mathbf{G}(V, E)$ Let $\lambda_1 \leq \cdots \leq \lambda_{|V|}$ be the eigenvalues of A. Assume A is diagonalizable. Then for any $l > 0$, the function $\delta(\beta, \omega_N)$ in Corollary 6 can be written

$$\sqrt{1 - \Phi^{-1}(\varepsilon(\beta, N)|V|\lambda_n^l \kappa(P^*(\omega_N)), (n-1)/2, 1/2)}. \tag{17}$$

Corollary 9 provides a probabilistic upper bound in Theorem 5, as a function of the largest eigenvalue of the adjacency matrix of $\mathbf{G}$. One can see that an automaton with a small largest eigenvalue allows for a better estimate of the CJSR, for a fixed number of samples N and length l.

4. CONCLUSION

In this work, we extended the scope of data-driven stability analysis of hybrid systems by generalizing previous data-driven results to the constrained case. In particular we have built our results on the basis of Berger et al. (2021).

We proceeded as follows. We first proposed a lifting result allowing us to reduce the computation of the CJSR of a given CSLS to the computation of a simpler JSR. We then stated the main theorem of this paper, which provides probabilistic guarantees for the stability of a given noise-free CSLS. Finally, we claimed that in case of uniformity on the distribution of switching sequences, we can investigate further the obtained bound. We showed that a smaller entropy of the automaton allows for a better guarantee about the stability.

In further research, we plan to extend this type of method to noisy observations. We also plan to investigate different approaches. For example, getting rid of the lifting result would allow to reduce the conservativism introduced by the latter, i.e. the gap between the lifted JSR $\rho(\mathbf{\Pi}_l)^{1/l}$ and the true CJSR $\rho(\mathbf{G}, \mathbf{\Sigma})$ in (9). In this regard we plan to directly approximate *multiple Lyapunov functions* (Philippe and Jungers, 2015, Definition 2).

REFERENCES

Berger, G.O., Jungers, R.M., and Wang, Z. (2021). Chance-constrained quasi-convex optimization with application to data-driven switched systems control. *arXiv:2101.01415 [cs, eess, math]*.

Calafiore, G. (2010). Random convex programs. *SIAM Journal on Optimization*, 20, 3427–3464. doi: 10.1137/090773490.

Campi, M., Lecchini, A., and Savaresi, S. (2003). An application of the virtual reference feedback tuning method to a benchmark problem. *European Journal of Control*, 9(1), 66–76. doi:https://doi.org/10.3166/ejc.9.66-76.

Dai, X. (2011). A gel'fand-type spectral radius formula and stability of linear constrained switching systems. *arXiv:1107.0124 [cs, math]*.

Hjalmarsson, H., Gevers, M., Gunnarsson, S., and Lequin, O. (1998). Iterative feedback tuning: theory and applications. *IEEE Control Systems Magazine*, 18(4), 26–41. doi:10.1109/37.710876.

Jungers, R. (2009). *The Joint Spectral Radius: Theory and Applications*. Springer.

Karimi, A. and Kammer, C. (2017). A data-driven approach to robust control of multivariable systems by convex optimization. *Automatica*, 85, 227–233. doi: 10.1016/j.automatica.2017.07.063.

Kenanian, J., Balkan, A., Jungers, R.M., and Tabuada, P. (2019). Data driven stability analysis of black-box switched linear systems. *Automatica*, 109.

Lin, H. and Antsaklis, P. (2009). Stability and stabilizability of switched linear systems: A survey of recent results. *Automatic Control, IEEE Transactions on*, 54, 308 – 322. doi:10.1109/TAC.2008.2012009.

Lind, D. and Marcus, B. (1995). *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press. doi:10.1017/CBO9780511626302.

Philippe, M., Essick, R., Dullerud, G.E., and Jungers, R.M. (2016). Stability of discrete-time switching systems with constrained switching sequences. *Automatica*, 72, 242–250. doi: https://doi.org/10.1016/j.automatica.2016.05.015.

Philippe, M. and Jungers, R.M. (2015). Converse lyapunov theorems for discrete-time linear switching systems with regular switching sequences. In *2015 European Control Conference (ECC)*, 1902–1907. doi: 10.1109/ECC.2015.7330816.

Rubbens, A., Wang, Z., and Jungers, R.M. (2021). Data-driven stability analysis of switched linear systems with sum of squares guarantees.

Tabuada, P. (2009). *Verification and Control of Hybrid Systems: A Symbolic Approach*. doi:10.1007/978-1-4419-0224-5.

Xu, X. and Acikmese, B. (2020). Approximation of the constrained joint spectral radius via algebraic lifting.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Realization formulas for involutions of matrix-valued positive real odd functions

S. TER HORST * **A. VAN DER MERWE** **

* *Department of Mathematics, Pure and Applied Analytics, North-West University, Potchefstroom, 2531 South Africa and DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) (Sanne.TerHorst@nwu.ac.za).*
** *Faculty of Engineering and the Built Environment, Academic Development Unit, University of the Witwatersrand, Johannesburg, 2000 South Africa and DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) (alma.vandermerwe@wits.ac.za).*

Abstract: In a seminal paper Foster (1924) showed that the impedances of lumped electrical circuits generated by inductances and capacitors are positive real odd functions ($\mathcal{PRO}$ for short). For multi-port electrical systems built from inductances and capacitors one obtains matrix-valued $\mathcal{PRO}$ functions, denoted $\mathcal{PRO}_m$ in the case of $m \times m$ matrix functions. Like $\mathcal{PRO}$, the class of matrix functions $\mathcal{PRO}_m$ is also a convex invertible cone, i.e., a convex cone closed under inversion (in the form of involution). Given a minimal, Weierstrass descriptor realization for a function in $\mathcal{PRO}_m$, we explicitly compute a minimal, Weierstrass descriptor realization for its involution, and through these formulas one can analyse the zero-pole structure of the function.

Keywords: Positive real odd matrix functions, convex invertible cones, descriptor systems, system inversion, transfer function zeros and poles.

1. INTRODUCTION

Positive real odd (rational) functions, often also referred to as positive real lossless functions, have been studied intensively in electrical engineering since it was observed by Foster (1924) that this class of functions, denoted $\mathcal{PRO}$, coincides with the impedances of lumped electrical circuits generated by inductances and capacitors. Foster also proved his famous canonical form for one-port reactance functions, namely, f is in $\mathcal{PRO}$ if and only if it has the form

$$f(z) = a_0 z + \sum_{k=1}^{s} \frac{a_k z}{z^2 + \omega_k^2}, \quad a_0 \geq 0,\ a_k, \omega_k \geq 0,\ 1 \leq k \leq s.$$

From this formula it is clear that all poles are simple, lie on in the imaginary axis $\mathrm{i}\mathbb{R}$ and have positive residue. Based on this formula, it is also possible to show that $\mathcal{PRO}$ is a convex invertible cone, cic for short, that is, a convex cone which is closed under inversion: For $f \in \mathcal{PRO}$, also $z \mapsto 1/f(z) \in \mathcal{PRO}$. See Cohen et al. (2007) for more on the cic-structure of $\mathcal{PRO}$. As a consequence of the fact that $\mathcal{PRO}$ is a cic, for any $0 \not\equiv f \in \mathcal{PRO}$, the poles of $1/f$ are also simple and on $\mathrm{i}\mathbb{R}$, so that both the poles and zeros of f are on $\mathrm{i}\mathbb{R}$ and they interlace.

Multi-port electrical systems build from inductances and capacitors correspond to positive real odd rational matrix functions, and they have also been studied intensively; cf., the classical monographs Newcomb (1966); Belevitch (1968); Anderson et al. (1973) as well as more recent work of Berger et al. (2014); Chu et al. (2008); Reis (2010), to name just a few. We write $\mathcal{PRO}_m$ for the class of positive real odd rational matrix functions of size $m \times m$, that is, an $m \times m$ rational matrix function F is in $\mathcal{PRO}_m$ in case

(i) $\mathrm{Re}(F(z)) \geq 0$ for all $\mathrm{Re}(z) > 0$;
(ii) $F(t) \in \mathbb{R}^{m \times m}$ for all $t \in \mathbb{R}$;
(iii) $-F(z) = F(-\overline{z})^*$ for z not a pole of F.

To the best of our knowledge, the cic structure of the class of matrix-valued $\mathcal{PRO}$ functions has not been studied in detail. It is straightforward to see that $\mathcal{PRO}_m$ is a convex cone, and also not difficult to prove that $\mathcal{PRO}_m$ is closed under inversion (in the form of involution). However, here we focus on explicit inversion formulas of functions in $\mathcal{PRO}_m$. In particular, we present two minimal realization formulas for functions F in $\mathcal{PRO}_m$, one of which is in Weierstrass descriptor form, and use these to explicitly compute minimal realization formulas of the same type for $F(z)^{-1}$. This gives another proof of the fact that $\mathcal{PRO}_m$ is a cic, but it also provides a way to analyse the zero and pole structure of functions in $\mathcal{PRO}_m$.

There is also an analogue of the foster canonical form for the matrix case. Any function $F \in \mathcal{PRO}_m$ can be written

$$F(z) = zQ + R + \sum_{j=1}^{s} \frac{1}{z^2 + \omega_j^2} (zQ_j + R_j), \qquad (1.1)$$

---

* **Mathematics Subject Classification (2010).** Primary 34A09; Secondary 93B50, 93B55.

**This work is based on research supported in part by the National Research Foundation of South Africa (Grant Numbers 118513 and 127364).

*** This is a resubmission of a full paper accepted for MTNS 2020, now shortened to an extended abstract for MTNS 2022.

where $\omega_j \in \mathbb{R}_+$ and $Q, R, Q_j, R_j \in \mathbb{R}^{m\times m}$ matrices so that

$$Q,\, Q_j \geq 0, \quad R = -R^T,\, R_j = -R_j^T,\ j = 1, \dots s. \quad (1.2)$$

This canonical form has also been extensively studied in the classical literature including the question how $\mathcal{PRO}_m$ functions of the form (1.1) can be represented by a multi-port electrical system of inductances and capacitors; cf., Newcomb (1966) and Anderson et al. (1973) and references given there. However, not every rational matrix function F of the form (1.1), with $\omega_j \in \mathbb{R}_+$ and Q, R, Q_j, R_j matrices satisfying (1.2) are in $\mathcal{PRO}_m$. For instance, it is easy to verify that the function

$$F(z) = \frac{1}{z^2+\omega^2} R, \quad \text{with } \omega \in \mathbb{R}_+,\, R = -R^T \neq 0,$$

is not in $\mathcal{PRO}_m$. We are not aware of any earlier source where necessary and sufficient conditions for a function F of the form (1.1) to be in $\mathcal{PRO}_m$ are presented, apart from our own work in Ter Horst et al. (2021). In the next theorem we present such conditions.

Theorem 1.1. An $m \times m$ rational matrix function F is in $\mathcal{PRO}_m$ if and only if F is of the form

$$F(z) = zQ + R + \sum_{j=1}^{s} \frac{1}{z^2+\omega_j^2} \left(zQ_j + R_j\right),$$

where $\omega_j \in \mathbb{R}_+$, $Q, R, Q_j, R_j \in \mathbb{R}^{m\times m}$ with $Q, Q_j \geq 0$ and R, R_j skew-symmetric so that

$$-\omega_j Q_j \leq \mathrm{i}R_j \leq \omega_j Q_j, \quad j = 1, \dots, s. \quad (1.3)$$

Note that since Q_j and R_j are real, for (1.3) to hold it suffices to verify one of the inequalities; indeed, because the conjugate of a positive semidefinite matrix is also positive semidefinite, $\omega_j Q_j + \mathrm{i}R_j \geq 0$ implies $\omega_j Q_j - \mathrm{i}R_j \geq 0$, and conversely.

Furthermore, we briefly discuss the zero and pole structure of functions in $\mathcal{PRO}_m$, using the minimal, Weierstrass descriptor realization for a function in $\mathcal{PRO}_m$ and its involution.

The results presented here require rather technical and sometimes long proofs, which can be found in Ter Horst et al. (2021).

2. REALIZATION FORMULAS FOR $\mathcal{PRO}_M$

In this section we provide some realization formulas for functions in $\mathcal{PRO}_m$ and define what we mean by pole and zero multiplicity. The formulas are not so much novel, but are mainly required for the analysis later on.

2.1 Realization formulas

By compiling various results in Reis (2010), specifically Proposition 7 and Theorem 8, together with basic state space manipulations, the following transfer function representation of $\mathcal{PRO}_m$ functions transpires.

Theorem 2.1. A function F is in $\mathcal{PRO}_m$ if and only if it admits a realization of the form

$$F(z) = zM + D + B^T(zI_n - A)^{-1}B, \quad (2.1)$$

for some integer $n \geq 0$, $M, D \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times m}$ and $A \in \mathbb{R}^{n\times n}$ where

$$\begin{aligned} &M \geq 0 \quad -A^T = A, \quad -D^T = D, \\ &\text{and the pair } (A, B) \text{ is controllable.} \end{aligned} \quad (2.2)$$

By a direct computation one can verify that Theorem 2.1 leads to the following descriptor realization characterization of $\mathcal{PRO}_m$ in Weierstrass form. See Dai (1989) and Kunkel et al. (2006) for more details on descriptor systems and the Weierstrass form.

Theorem 2.2. A function F is in $\mathcal{PRO}_m$ if and only if it admits a minimal descriptor realization of the form

$$F(z) = D^\circ + C^{\circ T}(zE^\circ - A^\circ)^{-1}B^\circ,$$

where we set $q = \operatorname{rank} M$ and factor $M = K^T K$ with $K \in \mathbb{R}^{q\times m}$, and where we set

$$A^\circ = \begin{bmatrix} A & 0 & 0 \\ 0 & I_q & 0 \\ 0 & 0 & I_q \end{bmatrix},\ E^\circ = \begin{bmatrix} I_n & 0 & 0 \\ 0 & 0 & I_q \\ 0 & 0 & 0 \end{bmatrix},\ B^\circ = \begin{bmatrix} B \\ 0 \\ -K \end{bmatrix},\ C^\circ = \begin{bmatrix} B \\ K \\ 0 \end{bmatrix}$$

and $D^\circ = D$, with $M, D \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{n\times m}$ and $A \in \mathbb{R}^{n\times n}$ matrices satisfying (2.2).

2.2 Poles and zeros

A number $z \in \mathbb{C} \cup \{\infty\}$ is a pole of $F \in \mathcal{PRO}_m$ simply when it is a pole of one of its entries. In the realization formulas of Theorems 2.1 and 2.2 the finite poles correspond to the eigenvalues of A. The multiplicity of a finite pole z of F is then defined as the dimension of the eigenspace of z as an eigenvalue of A, while the multiplicity of ∞ as a pole of F is defined as $\operatorname{rank} M$, provided $M \neq 0$. Since A is real and skew-symmetric, all poles are on $\mathrm{i}\mathbb{R}$ and the multiplicities of the finite poles add up to the McMillan degree of the proper part of F. Note that this definition of pole multiplicity is independent of the choice of the minimal realization.

Using the specific structure of the realizations obtained in Theorem 2.1 one can prove that the multiplicities cannot exceed the size of the matrix function.

Corollary 2.3. For $F \in \mathcal{PRO}_m$ every pole on $\mathrm{i}\mathbb{R}$, ∞ included, has a multiplicity of at most m.

In case $\det F(z) \not\equiv 0$ we say that F is invertible, with inverse given by the involution $F(z)^{-1}$, which is also in $\mathcal{PRO}_m$. In this case, we define the zeros of F to be the poles of $F(z)^{-1}$ and the zero-multiplicities of F are defined as the corresponding pole-multiplicities of $F(z)^{-1}$. The cic structure of $\mathcal{PRO}_m$ provides the following result.

Corollary 2.4. For $F \in \mathcal{PRO}_m$ every zero on $\mathrm{i}\mathbb{R}$, ∞ included, has a multiplicity of at most m.

3. INVERSION OF $\mathcal{PRO}_M$ FUNCTIONS

Let $F \in \mathcal{PRO}_m$ with $\det F(z) \not\equiv 0$. Then F is invertible, with inverse in $\mathcal{PRO}_m$ as well. In particular, the inverse of F has realization formulas as in Theorems 2.1 and 2.2. Throughout this section we assume $F \in \mathcal{PRO}_m$ is given in the state space realization form of Theorem 2.1. We express, in terms of the matrices in the realization (2.1)–(2.2), when $\det F(z) \not\equiv 0$, and in this case we present realization formulas of the types in Theorems 2.1 and 2.2 for the inverse of F.

3.1 Invertibility of $\mathcal{PRO}_m$ functions

By the inversion result for descriptor systems from Martins et al. (2007), together with a Schur complement computation, one obtains the following characterization for invertibility of F and of its inverse.

Proposition 3.1. Let $F \in \mathcal{PRO}_m$ be given by (2.1)-(2.2). Then for any $z \in \mathbb{C}$ we have

$$\det F(z) \not\equiv 0 \iff \det\left(\left[\begin{smallmatrix} zI_n & 0 \\ 0 & zM \end{smallmatrix}\right] - \left[\begin{smallmatrix} A & B \\ -B^T & -D \end{smallmatrix}\right]\right) \not\equiv 0.$$

Moreover, in that case we have

$$F(z)^{-1} = \left[\begin{smallmatrix} 0 & I_m \end{smallmatrix}\right]\left(\left[\begin{smallmatrix} zI_n & 0 \\ 0 & zM \end{smallmatrix}\right] - \left[\begin{smallmatrix} A & B \\ -B^T & -D \end{smallmatrix}\right]\right)^{-1}\left[\begin{smallmatrix} 0 \\ I_m \end{smallmatrix}\right].$$

Since $\left[\begin{smallmatrix} A & B \\ -B^T & -D \end{smallmatrix}\right]$ is also skew-symmetric, the invertibility criteria of Proposition 3.1 can be expressed in terms of a condition independent of the z variable.

Lemma 3.2. Let $F \in \mathcal{PRO}_m$ be given by (2.1)-(2.2). Then $\det F(z) \not\equiv 0$ if and only if $\operatorname{Ker}\left(\left[\begin{smallmatrix} B \\ D \end{smallmatrix}\right]|_{\operatorname{Ker} M}\right) = \{0\}$.

3.2 Minimal realizations of $F(z)^{-1}$

The realization in Proposition 3.1 will in general not be minimal, and hence some of the poles of the resolvent may not be poles of $F(z)^{-1}$, or the multiplicities may be inflated. To obtain a minimal realization, we decompose the matrices M, D and B with respect to the decomposition of $\mathbb{R}^m$ given by

$$\mathbb{R}^m = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \mathcal{X}_3, \tag{3.1}$$

with

$$\begin{aligned} \mathcal{X}_1 &= \operatorname{Ker} M^{\perp}, \\ \mathcal{X}_2 &= \operatorname{Ker}(P_{\operatorname{Ker} M} D|_{\operatorname{Ker} M})^{\perp}, \\ \mathcal{X}_3 &= \operatorname{Ker}(P_{\operatorname{Ker} M} D|_{\operatorname{Ker} M}), \end{aligned}$$

which yields decompositions of the form

$$B^T = \begin{bmatrix} B_1^T \\ B_2^T \\ B_3^T \end{bmatrix},\ D = \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ -D_{12}^T & D_{22} & 0 \\ -D_{13}^T & 0 & 0 \end{bmatrix},\ M = \begin{bmatrix} M_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.2}$$

with M_1 and D_{22} invertible. In particular, M_1 is positive definite and D_{22} is invertible and real, skew-symmetric, so that $\mathcal{X}_2$ must have even dimension. We set

$$m_1 = \dim \mathcal{X}_1, \quad m_2 = \dim \mathcal{X}_2, \quad m_3 = \dim \mathcal{X}_3,$$

so that $m = m_1 + m_2 + m_3$ and m_2 is even.

Furthermore, consider linear maps K_1 and Ξ so that

$$\begin{aligned} K_1 : \mathcal{X}_1 \to \mathbb{R}^{m_1}, \quad K_1^T K_1 = M_1, \\ \Xi : \mathcal{X}_3 \to \mathbb{R}^{m_3}, \quad \Xi^T \Xi = I_{\mathcal{X}_3}. \end{aligned} \tag{3.3}$$

Note that K_1 is invertible and Ξ orthogonal. Define

$$\begin{aligned} \widetilde{A} &= \begin{bmatrix} A - B_2 D_{22}^{-1} B_2^T & (B_1 + B_2 D_{22}^{-1} D_{12}^T) K_1^{-1} \\ K_1^{-T}(-B_1^T + D_{12} D_{22}^{-1} B_2^T) & -K_1^{-T}(D_{11} + D_{12} D_{22}^{-1} D_{12}^T) K_1^{-1} \end{bmatrix}, \\ \widetilde{B} &= \begin{bmatrix} B_3 \Xi^T \\ -K_1^{-T} D_{13} \Xi^T \end{bmatrix}. \end{aligned} \tag{3.4}$$

In terms of the decomposition (3.1), the condition for $\det F(z) \not\equiv 0$ is equivalent to $\operatorname{Ker}\left[\begin{smallmatrix} B_3 \\ D_{13} \end{smallmatrix}\right] = \{0\}$, or, equivalently, $\operatorname{Ker} \widetilde{B} = \{0\}$.

We are now ready to present the minimal Weierstrass realization for the inverse of F.

Theorem 3.3. Let $F \in \mathcal{PRO}_m$ be given by (2.1)-(2.2) and decompose B, D, M with respect to the decomposition (3.1) of $\mathbb{R}^m$ as in (3.2). Define $\widetilde{A}$ and $\widetilde{B}$ as in (3.4), with K_1 and Ξ as in (3.3), and assume $\operatorname{Ker} \widetilde{B} = \{0\}$ so that $\det F(z) \not\equiv 0$. Set $k = n + m_1 - m_3$ and let $\Gamma \in \mathbb{R}^{(n+m_1)\times k}$ be an isometry with $\operatorname{Im} \Gamma \perp \operatorname{Im} \widetilde{B}$. Then a minimal Weierstrass descriptor realization of the inverse of F is given by

$$F(z)^{-1} = D_{\text{inv}}^{\circ} + C_{\text{inv}}^{\circ T}(zE_{\text{inv}}^{\circ} - A_{\text{inv}}^{\circ})^{-1} B_{\text{inv}}^{\circ} \tag{3.5}$$

with

$$\begin{aligned} E_{\text{inv}}^{\circ} &= \begin{bmatrix} I_k & 0 & 0 \\ 0 & 0 & I_{m_3} \\ 0 & 0 & 0 \end{bmatrix}, \quad A_{\text{inv}}^{\circ} = \begin{bmatrix} A_{\text{inv}} & 0 & 0 \\ 0 & I_{m_3} & 0 \\ 0 & 0 & I_{m_3} \end{bmatrix}, \\ B_{\text{inv}}^{\circ} &= \begin{bmatrix} B_{\text{inv}} \\ 0 \\ -K_{\text{inv}} \end{bmatrix}, \quad C_{\text{inv}}^{\circ} = \begin{bmatrix} B_{\text{inv}} \\ K_{\text{inv}} \\ 0 \end{bmatrix}, \\ D_{\text{inv}}^{\circ} &= \begin{bmatrix} 0 & 0 & M_1^{-1} D_{13} \Phi_{33}^{-1} \\ 0 & D_{22}^{-1} & -D_{22}^{-1} \Phi_{23}^T \Phi_{33}^{-1} \\ -\Phi_{33}^{-1} D_{13}^T M_1^{-1} & \Phi_{33}^{-1} \Phi_{23} D_{22}^{-1} & \Phi_{33}^{-1} \Xi^T \widetilde{B}^T \widetilde{A} \widetilde{B} \Xi \Phi_{33}^{-1} \end{bmatrix}, \end{aligned} \tag{3.6}$$

where we define

$$A_{\text{inv}} = \Gamma^T \widetilde{A} \Gamma, \qquad K_{\text{inv}} = \begin{bmatrix} 0 & 0 & \Xi \Phi_{33}^{-1/2} \end{bmatrix}, \tag{3.7}$$

$$B_{\text{inv}} = \Gamma^T \begin{bmatrix} 0 & B_2 D_{22}^{-1} & (\Phi_{12} - B_2 D_{22}^{-1} \Phi_{23}) \Phi_{33}^{-1} \\ K_1^{-T} & -K_1^{-T} D_{12} D_{22}^{-1} & -K_1^{-T}(\Phi_{22} - D_{12} D_{22}^{-1} \Phi_{23}) \Phi_{33}^{-1} \end{bmatrix},$$

$$\Phi_{33} = B_3^T B_3 + D_{13}^T M_1^{-1} D_{13},$$

$$\Phi_{23} = B_2^T B_3 + D_{12}^T M_1^{-1} D_{13},$$

$$\Phi_{12} = A B_3 - B_1 M_1^{-1} D_{13}, \quad \Phi_{22} = B_1^T B_3 - D_{11} M_1^{-1} D_{13}$$

and where

$$\begin{aligned} \Xi^T \widetilde{B}^T \widetilde{A} \widetilde{B} \Xi = {} & B_3^T A B_3 - B_3^T B_1 M_1^{-1} D_{13} + D_{13}^T M_1^{-1} B_1^T B_3 \\ & - D_{13}^T M_1^{-1} D_{11} M_1^{-1} D_{13} - \Phi_{23} D_{22}^{-1} \Phi_{23}. \end{aligned}$$

Note that the descriptor realization for $F(z)^{-1}$ of Theorem 3.3 has precisely the form of the realization in Theorem 2.2. Reversing the argument in Section 2.1, we also obtain a realization of the type in Theorem 2.1.

Theorem 3.4. Let $F \in \mathcal{PRO}_m$ be given by (2.1)-(2.2) and decompose B, D, M with respect to the decomposition (3.1) of $\mathbb{R}^m$ as in (3.2). Assume $\det F(z) \not\equiv 0$. Then

$$F(z)^{-1} = zM_{\text{inv}} + D_{\text{inv}} + B_{\text{inv}}^T (zI_n - A_{\text{inv}})^{-1} B_{\text{inv}},$$

where B_{inv} and A_{inv} are as in (3.7), $D_{\text{inv}} = D_{\text{inv}}^{\circ}$ and $M_{\text{inv}} = K_{\text{inv}}^T K_{\text{inv}}$ with D_{inv}° as in (3.6) and K_{inv} as in (3.7). Moreover, the pair $(A_{\text{inv}}, B_{\text{inv}})$ is controllable.

Since Φ_{33} is invertible and Ξ an isometry, it is directly clear from the formula of M_{inv} that the pole multiplicity of $F(z)^{-1}$ at ∞ is equal to m_3, hence ∞ is a zero of F with multiplicity m_3. To say something about the zero multiplicities of F for finite zeros requires more analysis.

4. ZEROS AND POLES OF $\mathcal{PRO}_M$ FUNCTIONS

Recall that the zeros of F are defined as the poles of $F(z)^{-1}$. Hence, the finite zeros of F are given by the eigenvalues of A_{inv} with multiplicities equal to the dimensions of the corresponding eigenspaces. Thus, to understand the relation between zeros and poles one has to analyse the spectrum of A_{inv} in relation to the spectrum of A. There are three steps from A to A_{inv} that influence the eigenvalues:

(i) The extension of A into

$$\begin{bmatrix} A & B \\ -B^T & -D \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 & B_3 \\ -B_1^T & -D_{11} & -D_{12} & -D_{13} \\ -B_2^T & D_{12}^T & -D_{22} & 0 \\ -B_3^T & D_{13}^T & 0 & 0 \end{bmatrix}.$$

(ii) The transfer to $\widetilde{A}$ in (3.4) by taking the Schur complement with respect to $-D_{22}$.

(iii) The compression from $\widetilde{A}$ to $A_{\rm inv}$ in (3.7) via the isometry Γ.

The fact that all involved matrices are real skew-symmetric simplifies matters and enables us to use the Cauchy interlacing theorem, cf. Theorem 1 in Smith (1992), to arrive at the following result for the poles and zeros of functions in $\mathcal{PRO}_m$.

Theorem 4.1. Let $F \in \mathcal{PRO}_m$ be given by a minimal state space realization (2.1)-(2.2), so that F^{-1} has a minimal state space realization as in Theorem 3.4. Then for any integer $j \geq 0$ we have

$$\lambda_{j-\frac{m_2}{2}-m_3}(\mathrm{i}A_{\rm inv}) \leq \lambda_j(\mathrm{i}A) \leq \lambda_{j+1}(\mathrm{i}A) \leq \lambda_{j+\frac{m_2}{2}+m_1+1}(\mathrm{i}A_{\rm inv}),$$
$$\lambda_{j-\frac{m_2}{2}-m_1}(\mathrm{i}A) \leq \lambda_j(\mathrm{i}A_{\rm inv}) \leq \lambda_{j+1}(\mathrm{i}A_{\rm inv}) \leq \lambda_{j+\frac{m_2}{2}+m_3+1}(\mathrm{i}A).$$

In particular, if $0 \leq \omega_j < \omega_{j+1}$ are such that $\mathrm{i}\omega_j$ and $\mathrm{i}\omega_{j+1}$ are subsequent poles of F, then in the interval $(\mathrm{i}\omega_j, \mathrm{i}\omega_{j+1})$ on $\mathrm{i}\mathbb{R}$ F can have zeros whose multiplicities do not add up to more than m. Moreover, if $0 \leq \nu_j < \nu_{j+1}$ are such that $\mathrm{i}\nu_j$ and $\mathrm{i}\nu_{j+1}$ are subsequent zeros of F, then in the interval $(\mathrm{i}\nu_j, \mathrm{i}\nu_{j+1})$ on $\mathrm{i}\mathbb{R}$ F can have poles whose multiplicities do not add up to more than m.

Note that, unlike in the scalar case, for $m > 1$ it is possible that poles and zeros of $F \in \mathcal{PRO}_m$ occur at the same point on $\mathrm{i}\mathbb{R}$. Hence, as in the theorem, if $\mathrm{i}\omega_j$ and $\mathrm{i}\omega_{j+1}$ are subsequent poles of F, then zeros with a multiplicities adding up to at most m can occur between $\mathrm{i}\omega_j$ and $\mathrm{i}\omega_{j+1}$, but the theorem does not exclude the possibility that F also has zeros at $\mathrm{i}\omega_j$ and $\mathrm{i}\omega_{j+1}$.

5. CONCLUSION

We provided realization formulas for the inverses of functions in $\mathcal{PRO}_m$. These formulas enabled us to study the relations between zeros and poles of such functions. We also extended results on the Foster canonical form for functions in $\mathcal{PRO}_m$ by providing a necessary and sufficient condition for such a canonical form to give a function in $\mathcal{PRO}_m$ that does not seem to have appeared in the literature before.

We specifically focused on real-valued matrix functions, i.e., with $F(t) \in \mathbb{R}^{m\times m}$ for $t \in \mathbb{R}$, since without this condition we do not expect that so much can be said about the poles and zeros of such functions. However, we do expect that analogous results will exist for complex-valued lossless matrix functions regarding the realization formulas for their inverses as well as the Foster form. This may be a topic for future work. In this regard, it may also be interesting to investigate the case of reciprocal matrix functions in light of the results obtained here.

ACKNOWLEDGEMENTS

This work is based on research supported in part by the National Research Foundation of South Africa (NRF) and the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS). Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF and CoE-MaSS do not accept any liability in this regard.

REFERENCES

B.D.O. Anderson, A system theory criterion for positive real matrices, *SIAM J. Control* **5** (1967), 171–182.

B.D.O. Anderson, S. Vongpanitlerd, Networks Analysis and Synthesis, A Modern Systems Theory Approach, Prentice-Hall, New Jersey, 1973.

V. Belevitch, Classical Network Theory, Holden-Day, San-Francisco, 1968.

Thomas Berger and Timo Reis, Structural properties of positive real and reciprocal rational matrices, Proc. 21st MTNS, Groningen, The Netherlands, 2014.

D. Chu and R.C.E. Tan, *Algebraic characterizations for positive realness of descriptor systems SIAM J. Matrix Anal. Appl.* **30** (2008), 197—222.

N. Cohen and I. Lewkowicz, Convex invertible cones and positive real analytic functions, *Linear Algebra Appl.* **425** (2007), 797–813.

L. Dai, *Singular control systems*, Lecture Notes in Control and Information Sciences **118**, Springer-Verlag, Berlin, 1989.

R.M. Foster, A reactance theorem, *Bell System Technical Journal* **3** (1924), 259-–267.

R.A. Horn and C.R. Johnson, *Matrix Analysis,* Cambridge U.P., Cambridge, 1985.

S. ter Horst and A. Naudé, The convex invertible cone structure of positive real odd rational matrix functions, *Oper. Matrices* **15** (2021), 357–379.

P. Kunkel and V. Mehrmann, *Differential-algebraic equations, Analysis and numerical solution*, EMS Textbooks in Mathematics, European Mathematical Society (EMS), Zürich, 2006.

N. Martins, P.C. Pellanda, and J. Rommes, Computation of transfer function dominant zeros with applications to oscillation damping control of large power systems, *IEEE transactions on power systems* **22** (2007), 1657–1664.

R.W. Newcomb, *Linear multiport synthesis*, McGraw-Hill, 1966.

T. Reis, Circuit synthesis of passive descriptor systems—a modified nodal approach, *International Journal of Circuit Theory and Applications* **38.1** (2010), 44–68.

R.L. Smith, Some interlacing properties of the Schur complement of a Hermitian matrix, *Linear Algebra Applications* **177** (1992), 137—144.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The twofold Ellis-Gohberg inverse problem for rational matrix functions

S. ter Horst * **M.A. Kaashoek** ** **F. van Schagen** ***

* *Department of Mathematics, Pure and Applied Analytics, North-West University, Potchefstroom, 2531 South Africa, and DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) (e-mail: Sanne.TerHorst@nwu.ac.za).*
** *Department of Mathematics, VU Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands (e-mail: m.a.kaashoek@vu.nl)*
*** *Department of Mathematics, VU Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands (e-mail: f.van.schagen@vu.nl)*

Abstract: In recent years various papers appeared that concentrate on inverse problems associated with work of Ellis and Gohberg on orthogonal matrix Wiener functions. Here we study such an inverse problem restricting to rational matrix functions on the real line. The functions used in stating the problem are assumed to be given by minimal state space realizations, and the necessary and sufficient solution criterion as well as the formulas for the solution presented here are described in terms of the matrices of the state space realizations along with solutions to certain Lyapunov equations associated with the data.

Keywords: Inverse problem, rational matrix functions, state space realizations.

1. INTRODUCTION

The monograph by Ellis et al. (2003) collects and expands work of the authors from about two decades on systems of orthogonal matrix polynomials and matrix functions, in some instances together with D.C. Lay; see Ellis et al. (1992, 1995, 1996) for some of the original work, along with the monograph for additional references. This work also contains several results on related inverse problems, but much of the work on these inverse problems, specifically for the case of non-square matrix functions, only appeared recently, in Kaashoek et al. (2014, 2016, 2019) and Ter Horst et al. (2017, 2019, 2020). Here we present some of the work that will mainly appear in Ter Horst et al. (2020).

2. FORMULATION OF THE PROBLEM

The inverse problem we consider here has a data set which consists of four proper rational matrix functions $\alpha, \beta, \gamma, \delta$ with sizes given by

$$\begin{aligned} &\alpha(\lambda) \in \mathbb{C}^{p\times p}, \quad \beta(\lambda) \in \mathbb{C}^{p\times q}, \\ &\gamma(\lambda) \in \mathbb{C}^{q\times p}, \quad \delta(\lambda) \in \mathbb{C}^{q\times q}, \end{aligned} \tag{2.1}$$

and where α and β have only poles in the open lower half plane $\mathbb{C}_-$, γ and δ have poles only in the open upper half plane $\mathbb{C}_+$, and with values at ∞ given by

$$\alpha(\infty) = I_p, \quad \beta(\infty) = 0, \quad \gamma(\infty) = 0, \quad \delta(\infty) = I_q.$$

When the above properties are fulfilled we call $\{\alpha, \beta, \gamma, \delta\}$ an *admissible rational data set.*

⋆ This work is based on research supported in part by the National Research Foundation of South Africa (Grant Numbers 118513 and 127364).
This is a resubmission of an extended abstract that was accepted for the 2020 MTNS conference.

Given an admissible rational data set $\{\alpha, \beta, \gamma, \delta\}$ the rational version of the twofold Ellis-Gohberg inverse problem is to find a strictly proper $p \times q$ rational matrix function g which has all its poles in $\mathbb{C}_-$ such that

$$\begin{aligned} &\alpha(\lambda) + g(\lambda)\gamma(\lambda) - I_p \text{ has poles only in } \mathbb{C}_+; \\ &g^*(\lambda)\alpha(\lambda) + \gamma(\lambda) \text{ has poles only in } \mathbb{C}_-; \\ &\delta(\lambda) + g^*(\lambda)\beta(\lambda) - I_q \text{ has poles only in } \mathbb{C}_-; \\ &g(\lambda)\delta(\lambda) + \beta(\lambda) \text{ has poles only in } \mathbb{C}_+. \end{aligned} \tag{2.2}$$

Here and in the remainder, for any rational matrix function $\varphi(\lambda)$ the *adjoint* of the function $\varphi(\lambda)$ indicates function $\varphi(\overline{\lambda})^*$ and will be denoted by $\varphi^*(\lambda)$. We shall refer to the above problem as the *twofold Rat-EG inverse problem.* Also in other contexts, EG will be used to abbreviate Ellis-Gohberg.

We briefly mention the direct EG-problem from which the inverse EG problem studied here is derived. The starting point is a strictly proper $p \times q$ rational matrix function g which has all its poles in $\mathbb{C}_-$, and the problem is to obtain an admissible rational data set $\{\alpha, \beta, \gamma, \delta\}$ such that (2.2) is satisfied. Solving this problem is equivalent to solving two linear equations, $Tx_1 = f_1$ and $Tx_2 = f_2$, where $T : \mathcal{X} \to \mathcal{X}$ is a structured linear operator defined by the given function g, and the right hand sides f_1 and f_2 are specific elements in $\mathcal{X}$ defined by g too; see formula (1.9) in Ter Horst et al. (2019) for more details. The solutions x_1 and x_2, if they exist, will uniquely determine the requested data set $\{\alpha, \beta, \gamma, \delta\}$. Moreover, under an additional simple finite dimensional condition, the inverse of T can be expressed in terms of the data set $\{\alpha, \beta, \gamma, \delta\}$.

3. OPERATOR THEORY SOLUTION

This version was investigated in the more general context of matrix Wiener class functions on the real line in Ter Horst et al. (2019), where necessary and sufficient conditions for the existence of solutions were obtained. Also, in case a solution exists, it is shown that this is the unique solution and two formulas for the solution are presented. Based on the results in Ter Horst et al. (2019), see Ter Horst et al. (2020) for details, it is easily verified that for the rational matrix data functions $\{\alpha, \beta, \gamma, \delta\}$ of an admissible rational data set, necessary conditions are:

$$\begin{aligned} &\text{(C1)} \quad \alpha^*(\lambda)\alpha(\lambda) - \gamma^*(\lambda)\gamma(\lambda) = I_p; \\ &\text{(C2)} \quad \delta^*(\lambda)\delta(\lambda) - \beta^*(\lambda)\beta(\lambda) = I_q; \\ &\text{(C3)} \quad \alpha^*(\lambda)\beta(\lambda) = \gamma^*(\lambda)\delta(\lambda). \end{aligned}$$

In addition to these three equations, in the Wiener class function inverse problem on the real line of Ter Horst et al. (2019), two more conditions are required to obtain necessary and sufficient conditions, and these conditions can be formulated as the required that two operators, determined by Hankel operators associated with the data functions, are one-to-one; see Theorem A.1 in Ter Horst et al. (2020). The main aim of the present research is to express the two operator conditions in a more computationally effective way, and use the new versions of these conditions to obtain new representations of the solutions.

4. STATE SPACE PRELIMINARIES

To obtain more computationally attractive solution criteria and a more explicit description of the solution, we assume our data functions are given in the form of finite dimensional state space realizations:

$$\begin{aligned} \alpha(\lambda) &= I_p + iC_1(\lambda I_{n_1} - iA_1)^{-1}B_1, \\ \beta(\lambda) &= iC_2(\lambda I_{n_2} - iA_2)^{-1}B_2, \\ \gamma(\lambda) &= -iC_3(\lambda I_{n_3} + iA_3)^{-1}B_3, \\ \delta(\lambda) &= I_q - iC_4(\lambda I_{n_4} + iA_4)^{-1}B_4. \end{aligned} \tag{4.3}$$

Here A_j, $1 \leq j \leq 4$, is a square matrix which is assumed to be stable, e.g., all eigenvalues of A_j are in the open left half plane $\mathbb{C}_{\text{left}}$. These stability conditions are automatically fulfilled if the realizations are minimal. In the latter case the McMillan degrees of $\alpha, \beta, \gamma, \delta$ are equal to n_1, n_2, n_3, n_4, respectively. Although the functions $\alpha, \beta, \gamma, \delta$ in an admissible rational data set can always be represented in this way, we shall not require the realizations in (4.3) to be minimal.

To state our solution to the twofold Rat-EG inverse problem we shall use the solution P_{ij}, for $i, j \in \{1, 2\}$ or $i, j \in \{3, 4\}$, to the following Lyapunov equation associated with the pairs (A_i, C_i) and (A_j, C_j):

$$\begin{aligned} &A_i^* P_{ij} + P_{ij} A_j + C_i^* C_j = 0, \\ &\text{with } i, j \in \{1, 2\} \text{ or } i, j \in \{3, 4\}. \end{aligned} \tag{4.4}$$

For $i = j$ we abbreviate P_{jj} to P_j. We also need the solution Q_j, for $1 \leq j \leq 4$, to the Lyapunov equation

$$A_j Q_j + Q_j A_j^* + B_j B_j^* = 0, \quad 1 \leq j \leq 4. \tag{4.5}$$

Since the matrices A_j, $1 \leq j \leq 4$, are all stable, the solutions P_{ij} and Q_j to the Lyapunov equations (4.4) and (4.5) are unique, and given explicitly by

$$P_{ij} = \int_0^\infty e^{sA_i^*} C_i^* C_j e^{sA_j}\, ds, \; Q_j = \int_0^\infty e^{sA_j} B_j B_j^* e^{sA_j^*}\, ds.$$

From the latter identities it follows that the matrices $P_j = P_{jj}$ and Q_j, $1 \leq j \leq 4$, are nonnegative. Furthermore, we have $P_{12}^* = P_{21}$ and $P_{34}^* = P_{43}$. See Section 3.8 in Zhou et att. (1996) for the basic theory of Lyapunov equations; see also Theorem I.5.5 in Gohberg et al. (1990).

5. SOLUTION IN TERMS OF STATE SPACE REALIZATIONS

Since P_2 and Q_2 are nonnegative, the matrix $I_{n_2} + Q_2 P_2$ is invertible. Similarly, $I_{n_3} + Q_3 P_3$ is invertible because P_3 and Q_3 are nonnegative. Using the matrices defined above, we set

$$\begin{aligned} N_1 &:= P_1 - P_{12}(I_{n_2} + Q_2 P_2)^{-1} Q_2 P_{21}, \\ N_4 &:= P_4 - P_{43}(I_{n_3} + Q_3 P_3)^{-1} Q_3 P_{34}. \end{aligned} \tag{5.6}$$

Now we are ready to formulate our first main result, which provides necessary and sufficient conditions for the twofold Rat-EG inverse problem.

Theorem 1. The twofold Rat-EG inverse problem associated with the rational data set $\{\alpha, \beta, \gamma, \delta\}$ given by state space realizations (4.3) has a solution if and only if the following conditions are satisfied:

$$\begin{aligned} &\text{(R1)} \quad (C_1 + B_1^* P_1)(\lambda I_{n_1} - iA_1)^{-1} B_1 = \\ &\qquad = B_3^*(\lambda I_{n_3} - iA_3^*)^{-1} P_3 B_3; \\ &\text{(R2)} \quad (C_4 + B_4^* P_4)(\lambda I_{n_4} + iA_4)^{-1} B_4 = \\ &\qquad = B_2^*(\lambda I_{n_2} + iA_2^*)^{-1} P_2 B_2; \\ &\text{(R3a)} \quad B_1^*(\lambda I_{n_1} + iA_1^*)^{-1} P_{12} B_2 = B_3^* P_{34}(\lambda I_{n_4} + iA_4)^{-1} B_4; \\ &\text{(R3b)} \quad (C_2 + B_1^* P_{12})(\lambda I_{n_2} - iA_2)^{-1} B_2 = \\ &\qquad = B_3^*(\lambda I_{n_3} - iA_3^*)^{-1}(C_3^* + P_{34} B_4) \\ &\text{(R4)} \quad \text{the matrices } I_{n_1} - Q_1 N_1 \text{ and } I_{n_4} - Q_4 N_4 \text{ are invertible.} \end{aligned}$$

Moreover, in that case the solution is unique.

In case the necessary and sufficient conditions of the previous theorem are satisfied, it is possible to describe the unique solution g as in the following result.

Theorem 2. Assume the functions of the rational data set $\{\alpha, \beta, \gamma, \delta\}$ are given by state space realizations (4.3), and assume the conditions (R1)-(R4) are satisfied. Then the unique solution g to the twofold Rat-EG inverse problem is given by

$$\begin{aligned} g(\lambda) = &-iC_1(\lambda I_{n_1} - iA_1)^{-1} Y_1 + \\ &- iC_2(\lambda I_{n_2} - iA_2)^{-1} \left(Y_2 - \widetilde{Y_2} \right). \end{aligned}$$

Here Y_2 and $\widetilde{Y_2}$ are matrices of size $n_2 \times q$, and Y_1 is a matrix of size $n_1 \times q$, and these three matrices are defined by

$$\begin{aligned} Y_1 &= (I_{n_1} - Q_1 N_1)^{-1} Q_1 P_{12}(I_{n_2} + Q_2 P_2)^{-1} B_2, \\ Y_2 &= (I_{n_2} + Q_2 P_2)^{-1} B_2, \\ \widetilde{Y_2} &= (I_{n_2} + Q_2 P_2)^{-1} Q_2 P_{21} Y_1. \end{aligned}$$

The results of Ter Horst et al. (2019) also provide an alternative formula for the unique solution, which can be presented in terms of the state space formulas for the data functions and the associated matrices as in the next theorem.

Theorem 3. Assume the functions of the rational data set $\{\alpha, \beta, \gamma, \delta\}$ are given by state space realizations (4.3), and

assume the conditions (R1)-(R4) are satisfied. Then the unique solution g to the twofold Rat-EG inverse problem is given by

$$g(\lambda) = -iX_1(\lambda I_{n_4} - iA_4^*)^{-1}C_4^* + \\ - i\left(X_2 - \widetilde{X_2}\right)(\lambda I_{n_3} - iA_3^*)^{-1}C_3^*.$$

In this case X_2 and $\widetilde{X_2}$ are matrices of size $p \times n_3$, and X_1 is a matrix of size $p \times n_4$, and these three matrices are defined by

$$X_1 = B_3^*(I_{n_3} + P_3Q_3)^{-1}P_{34}Q_4(I_{n_4} - N_4Q_4)^{-1},$$
$$X_2 = B_3^*(I_{n_3} + P_3Q_3)^{-1},$$
$$\widetilde{X_2} = X_1P_{43}Q_3(I_{n_3} + P_3Q_3)^{-1}.$$

ACKNOWLEDGEMENTS

This work is based on research supported in part by the National Research Foundation of South Africa (NRF) and the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS). Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF and CoE-MaSS do not accept any liability in this regard.

REFERENCES

R.L. Ellis and I. Gohberg, Orthogonal systems related to infinite Hankel matrices, *J. Funct. Analysis* **109** (1992), 155–198.

R.L. Ellis and I. Gohberg, *Orthogonal systems and convolution operators*, Oper. Theory Adv. Appl. **140**, Birkhäuser Verlag, Basel, 2003.

R.L. Ellis, I. Gohberg, and D.C. Lay, Infinite analogues of block Toeplitz matrices and related orthogonal functions. *Integral Equ. Oper. Theory* **22** (1995), 375–419.

R. L. Ellis, I. Gohberg, and D.C. Lay, On a Class of Block Toeplitz Matrices, *Linear Algebra Appl.* **241/243** (1996), 225–245.

I. Gohberg, S. Goldberg, and M.A. Kaashoek, *Classes of Linear Operators I*, Oper. Theory Adv. Appl. **49**, Birkhäuser Verlag, Basel, 1990.

S. ter Horst, M.A. Kaashoek, and F. van Schagen, The discrete twofold Ellis-Gohberg inverse problem. *J. Math. Anal. Appl.* **452** (2017), 846—870.

S. ter Horst, M.A. Kaashoek, and F. van Schagen, The twofold Ellis-Gohberg inverse problem in an abstract setting and applications, in: *Interpolation and realization theory with applications to control theory*, Oper. Theory Adv. Appl. **272** (2019), 155–212.

S. ter Horst, M.A. Kaashoek, and F. van Schagen, The twofold Ellis-Gohberg inverse problem for rational matrix functions on the real line, *Oper. Theory Adv. Appl.* **279** (2020), 145–173.

M.A. Kaashoek and F. van Schagen, The inverse problem for Ellis-Gohberg orthogonal matrix functions, *Integral Equ. Oper. Theory* **80** (2014), 527—555.

M.A. Kaashoek and F. van Schagen, The Ellis-Gohberg inverse problem for matrix-valued Wiener functions on the line, *Oper. Matrices* **10** (2016), 1009–1042.

M.A. Kaashoek and F. van Schagen, Onefold and twofold Ellis-Gohberg inverse problems for scalar Wiener class functions, in: *Positivity and Noncommutative Analysis, Festschrift in Honour of Ben de Pagter on the Occasion of his 65th Birthday*, Birkhäuser Verlag, Basel, 2019.

K. Zhou with J. C. Doyle and K. Glover, *Robust and optimal control*, Prentice Hall, New Jersey, 1996.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On the Risks of Feedback-Trading

Michael Heinrich Baumann *

* *University of Bayreuth, 95447 Bayreuth, Germany (e-mail: michael.baumann@uni-bayreuth.de).*

Abstract: It has been shown in the literature that for certain trading strategies based on control techniques, namely for the so-called simultaneously long short strategies under relatively weak market assumptions in continuous time, the so-called robust positive expectation property holds. This means that for such strategies, if the assumptions are fulfilled, in expectation positive profits can be proven. Of course, arguments such as trading costs or trading constraints can be used when discussing these unexpected results. But there are also risks inherent in the strategies themselves, such as short-selling risks, discretization risks, or momenta. In this talk, we will present these risks and show how they can possibly be controlled.

Keywords: Feedback Trading, Financial Mathematics, Risk Measures, Skewness, Control-based Trading Strategies, Discretization, Simultaneously Long Short Trading, Stochastic Processes

EXTENDED ABSTRACT

Trading with stocks or, in general, assets means buying and selling them according to certain criteria. For this purpose, there is a wide variety of investment strategies. The strategy we will look at in this paper, the Simultaneously Long Short (SLS) strategy, belongs to a class of strategies called feedback strategies. These strategies are inspired and often analyzed by methods originating from control theory. As the name SLS suggests, this strategy invests both long and short, shifting more investment to the better performing side. Short selling means to sell assets that one does not own but only borrows. In such a negative investment profits are obtained for falling prices and losses for rising prices.

If we consider a single-asset market with price $p(t) \geq 0$ with $t = 0, h, 2h, \ldots, T$ with $h > 0$ or with $t \in [0, T]$ and denote by $I(t)$ the investment at time t, i.e. the number of shares held ($\nu(t)$) times their price per unit ($p(t)$), we can calculate the cumulative gain/loss via $g(t) = \sum_{i=1}^{t} \nu((i-1)h)(p(ih) - p((i-1)h)) = \sum_{i=1}^{t} I((i-1)h)\left(\frac{p(ih)}{p((i-1)h)} - 1\right)$ resp. $g(t) = \int_0^t \frac{I(t)}{p(t)} dp(t)$. The idea behind feedback strategies is that not only the profit is a function of the investment, but also the investment is a function of the profit, i.e., $I(t) = f(g(t))$ for an appropriate function f. Two possible choices for feedback strategies are the linear feedback strategies I^L and I^S with $f_L(x) = I_L^* + K_L \cdot x$ with $I_L^*, K_L > 0$ and $f_S(x) = I_S^* + K_S \cdot x$ with $I_S^*, K_S < 0$. The SLS strategy is defined as follows: $I^{SLS} = I^L + I^S$, where $I^* := I_L^* = -I_S^* > 0$ and $K := K_L = -K_S > 0$. Observe that when the time axis and the price process are continuous, the strategy I^L always invests long and the strategy I^S always invests short under the same conditions. Note that it is important that the gains of the long and the short side (g_L, g_S) are calculated separately (since otherwise gain and investment of the SLS rule would always be zero).

Performance properties of the SLS strategy are analyzed in various papers. Expected gains are calculated for prices governed by geometric Brownian motions (Barmish and Primbs, 2011, 2016; Dokuchaev, 2012; Dokuchaev and Savkin, 1998), by Merton's jump diffusion model (Baumann, 2017), by tree models (Iwarere and Barmish, 2014), in discrete time models with constant trend (Malekpour and Barmish, 2016; Baumann and Grüne, 2017), as well as in models with variable trends, so-called time-varying geometric Brownian motions (Primbs and Barmish, 2013) and discrete time models with variable parameters (Baumann, 2021), and also in other models. [1] The discrete time models are of particular interest since they try to avoid the joint hypotheses problem (i.e. the problem of simultaneously assuming a market model and the market efficiency hypothesis). Variances of gains of SLS strategies are calculated, e.g., for geometric Brownian motions (Barmish and Primbs, 2011), Merton's jump diffusion model (Baumann, 2017), discrete time models with constant and varying trends (Baumann, 2021), etc. The most interesting result concerning SLS trading is that under certain assumptions, in most of the mentioned models under almost all parameter settings the expected gain is positive. An exception are discrete time models with varying trends (Baumann, 2021), which gives a first hint on possible risks traders using the SLS rule—or, generally, feedback strategies—face. In the following, we explain four risks, some obvious, some hidden. And we show which questions might be quite interesting for future research.

Unreasonable assumptions

In theoretical results often assumptions are made, which might not be fulfilled in real-world. For example, trading costs might be neglected or it may be assumed that traders do not influence the price. There might be neither trading delay nor restrictions. All amounts of assets (long

[1] Please note that the cited literature represents only a small part of the available literature.

and short) may be available, and so on. Clearly, these assumptions must not hold true in real world. One solution would be to provide trading statistics with historical price data. However, then the researcher faces the problem of overfitting (Bailey et al., 2014; Wilmott, 2000). Also, p-hacking has to be kept in mind.

Short selling

While traders cannot lose more money on long investments than they invest, the possible loss on short sales—no matter how small—is in principle unbounded. Even if only a small amount is sold short, if the stock rises sharply, for example, because the underlying company receives a takeover bid, the investor may incur a very large loss. This is the reason why private persons or companies without very large financial reserves tend to avoid short sales. However, in order to profit from falling prices resp. to bet on falling prices, there are derivatives that have a similar payoff profile, but whose worst case risk can be limited. Infinitely large worst-case losses create a serious risk. It may be possible to circumvent or hedge this risk by means of derivatives such as put options—but how does such a hedging strategy influence the performance of the SLS rule?

Skewness and other higher momenta

Generally, it is rather hard or even impossible to calculate any classical risk measures for the SLS strategy. An exception is the case of geometric Brownian motions since there the density of the gain/loss distribution is known (Barmish and Primbs, 2011). A way to include risk in the performance investigations of SLS trading might be to calculate the skewness, which could be done with the same methodology and under similar assumptions as for the variance (Baumann, 2021). However, there is a fundamental problem: it is not clear whether higher or lower skewed gains are more risky/preferable. Highly skewed gains mean that there is a small chance to make high gains and a high risk to make small losses (as typically when thinking about lottery or betting slips); low skewed gains mean there is a small probability for a high loss and a high chance for small gains.

For example, consider the random variables X, Y modeling profits (defined on appropriate spaces). We assume that X is -1 with probability 99% and 99 with 1% and Y that is -99 with probability 1% and 1 with 99%. It holds: $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ and $Var(X) = Var(Y) = 99$. Clearly, X has a high skewness ($v(X) \approx 9,85$) while Y has a small skewness ($v(Y) \approx -9,85$).

When considering the probability of loosing ($P(X < 0) = 99\%$, $P(Y < 0) = 1\%$), Y is preferable. When considering the worst case, X is preferable. Next, we have look at a classical risk measure, the Value at Risk (Föllmer and Schied, 2011, Sec. 4.4). It holds $V@R_{1\%}(X) = 1$ and $V@R_{1\%}(Y) = -1$, thus, one would prefer Y, but $V@R_{0.5\%}(X) = 1$ and $V@R_{0.5\%}(Y) = 99$, speaking for X.

Discretization risks

The presumably most important risk is the discretization risk. Baumann (2021) shows that in discrete time (and all real trades take place in discrete time) traders face the risk of switching trends. While sign switching trends are not any problem for the continuous time SLS, they are in discrete time. For all two points in time where the sign of the trend is either positive or negative, the trader may expect a profit. Every time the trend switches its sign, the trader faces an expected loss. One way out would be to modify the strategy by model assumptions or by estimators, but this would be against the fundamental idea of feedback trading.

CONCLUSION

While for the SLS strategy one can show performance characteristics such as positive expected returns, both in theory and based on backtesting, its risks have not yet been studied in sufficient detail. In this work, we have presented the presumably most important risks and outlined against which risks one may or may not protect oneself. Further research on these topics, based on theory and backtesting, is desirable.

REFERENCES

B. R. Barmish and J. A. Primbs. On arbitrage possibilities via linear feedback in an idealized Brownian motion stock market. *IEEE Conference on Decision and Control and European Control Conference (CDC-ECC),* pages 2889–2894, 2011.

B. R. Barmish and J. A. Primbs. On a new paradigm for stock trading via a model-free feedback controller. *IEEE Transactions on Automatic Control,* 61:3, pages 662–676, 2016.

N. G. Dokuchaev. Dynamic portfolio strategies: quantitative methods and empirical rules for incomplete information. *Springer Science & Business Media,* 47, 2012.

N. G. Dokuchaev and A. V. Savkin. A hedging investment strategy with a positive average gain without market estimation. *2nd IMACS International Multiconference, Computational Engineering in Systems Applications,* pages 94–99, 1998.

M. H. Baumann. On stock trading via feedback control when underlying stock returns are discontinuous. *IEEE Transactions on Automatic Control,* 62:6, pages 2987–2992, 2017.

S. Iwarere and B. R. Barmish. On stock trading over a lattice via linear feedback. *IFAC Proceedings,* 47:3, pages 7799–7804, 2014.

S. Malekpour and B. R. Barmish. On stock trading using a controller with delay: the robust positive expectation property. *IEEE Decision and Control (CDC),* pages 2881–2887, 2016.

M. H. Baumann and L. Grüne. Simultaneously long short trading in discrete and continuous time. *Systems & Control Letters,* 99, pages 85–89, 2017.

J. A. Primbs and B. R. Barmish. On stock trading: Can a trend follower expect to win? *Midwest Finance Association Conference,* 2013.

M. H. Baumann. Beating the market? A mathematical puzzle for market efficiency. *Decisions in Economics and Finance,* 2021.

D. H. Bailey, J. Borwein, M. Lopez de Prado, and Q. J. Zhu. Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Notices Of The American Mathematical Society,* 61:5, pages 458–471, 2014.

P. Wilmott. The use, misuse and abuse of mathematics in finance. *Philosophical Transactions of the Royal Society of London, The Royal Society,* A 358, pages 63–73, 2000.

H. Föllmer and A. Schied. Stochastic finance. De Gruyter, 3rd edition, 2011.

ACKNOWLEDGEMENT

The author wishes to thank Lars Grüne (Universität Bayreuth), Bob Ross Barmish (Boston University), Michaela Baumann (Nürnberger Versicherung), and Bernhard Herz (Universität Bayreuth).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Spectrahedral Shadows and Completely Positive Maps on Real Closed Fields

Mario Kummer *

* *Technische Universität Dresden, Bereich Mathematik und Naturwissenschaften, Fakultät Mathematik, Institut für Geometrie (e-mail: mario.kummer@tu-dresden.de).*

Abstract: We provide a new characterization of spectrahedral shadows. We use this to show that the cone of copositive matrices of size at least five is not a spectrahedral shadow, a convex semialgebraic set that resisted the efforts of Scheiderer (2018). This is a joint work with Manuel Bodirsky and Andreas Thom.

Keywords: Spectrahedral shadows, real closed fields, completely positive maps, sums of squares. AMS subject classification: 11E25, 12D15, 90C22.

1. INTRODUCTION

Semidefinite programming is a generalization of linear programming whose feasible sets are called *spectrahedral shadows*. These are convex semialgebraic sets that (are the image under an affine linear map of a set that) can be described by symmetric linear matrix inequalities. Nemirovski (2007) asked whether every convex semialgebraic set is a spectrahedral shadow. Later Helton and Nie (2009) conjectured that the answer to this question is in fact *yes*. This conjecture was recently disproved by Scheiderer (2018). Further counter-examples were subsequently given by Fawzi (2021) and Bettiol et al. (2021). However, the techniques used in these articles were essentially the same as the ones developed by Scheiderer (2018). In a joint work with Manuel Bodirsky and Andreas Thom we provide new techniques for proving that a certain semialgebraic set is not a spectrahedral shadow. We use these to prove that the set of all *copositive matrices* of size m, i.e. the cone of all symmetric $m \times m$ matrices A such that $x^t Ax \geq 0$ for all $x \in \mathbb{R}^m_{\geq 0}$, is not a spectrahedral shadow whenever $m \geq 5$. In the following, we describe our techniques in more detail.

2. PRIMITIVE DEFINABILIY

Let X be a set and $R_i \subset X^{r_i}$, $i \in I$, a collection of relations over X. A first-order formula over the *relational structure* $\Gamma = (X; \{R_i : i \in I\})$ is *primitive positive* if it is of the form

$$\exists x_1, \ldots, x_n (\psi_1 \wedge \cdots \wedge \psi_m)$$

where $\psi_1, \ldots, \psi_m$ are atomic formulas formed with relations R_i, $i \in I$. In particular, no negation, disjunction, and universal quantification is allowed. An *endomorphism* of the relational structure Γ is a function $f : X \to X$ such that $R_i(x_1, \ldots, x_{r_i})$ implies $R_i(f(x_1), \ldots, f(x_{r_i}))$ for all $x_1, \ldots, x_{r_i} \in X$ and $i \in I$. An easy but important observation is that if R is a relation over X that can be defined using a primitive positive formula over Γ, then R is preserved by all endomorphisms of Γ. In many cases, the converse holds true as well. This was for instance shown by Bodnarčuk et al. (1969) in the case when X is finite.

3. A CHARACTERIZATION OF SPECTRAHEDRAL SHADOWS

Now let $\mathbf{R}$ be a real closed extension of $\mathbb{R}$. We consider the relational structure on $\mathbf{R}$ that is given by all spectrahedral shadows $R \subset \mathbf{R}^m$ which are defined by matrices with entries in $\mathbb{R}$. A function $f : \mathbf{R} \to \mathbf{R}$ is an endomorphism of this relational structure if and only if f is $\mathbb{R}$-linear, unital and completely positive in the sense that applying f entry-wise to a positive semidefinite matrix over $\mathbf{R}$ gives a positive semidefinite matrix. We use this observation to prove the following abstract characterization of spectrahedral shadows.

Theorem 1. Let $S \subset \mathbb{R}^n$ be a semi-algebraic set. Then S is a spectrahedral shadow if and only if for all real closed field extensions $\mathbf{R}$ of $\mathbb{R}$ and all $\mathbb{R}$-linear, unital and completely positive functions $f : \mathbf{R} \to \mathbf{R}$ the base-change $S(\mathbf{R}) \subset \mathbf{R}^n$ is preserved by component-wise application of f.

Here the base-change $S(\mathbf{R})$ of a semialgebraic set $S \subset \mathbb{R}^n$ is the semialgebraic subset of $\mathbf{R}^n$ defined by the same inequalities as S.

4. NEW COUNTER-EXAMPLES TO THE HELTON–NIE CONJECTURE

Relating Theorem 1 to sums of squares in the group ring of the abelian group $\mathbb{Q}^n$, we are able to prove the following criterion for the convex hull of the positive part of a toric variety to be a spectrahedral shadow.

Theorem 2. Let $S \subset \mathbb{Z}^m_{\geq 0}$ be a finite set of cardinality n. Let $X \subset \mathbb{R}^n$ be the image of $\mathbb{R}^m_{\geq 0}$ under the monomial map

$$\mathbb{R}^m \to \mathbb{R}^n,\ x \mapsto (x^\alpha : \alpha \in S).$$

The closed convex hull of X is a spectrahedral shadow if and only if for every real closed field $\mathbf{R}$ and every nonnegative polynomial $P \in \mathbf{R}[x_1, \ldots, x_m]$ with support

in S there is a $d \in \mathbb{Z}_{>0}$ such that $P(x_1^d, \ldots, x_m^d)$ is a sum of squares of polynomials.

The cone of *completely positive matrices* is the closed convex hull of the image of $\mathbb{R}^m_{\geq 0}$ under the monomial map given by all monomials of degree 2. We apply Theorem 2 to the *Horn polynomial*

$$(x_1+x_2+x_3+x_4+x_5)^2-4\cdot(x_1x_2+x_2x_3+x_3x_4+x_4x_5+x_5x_1)$$

that was studied by Hall jun. and Newman (1963). Like this, we prove that the cone of completely positive matrices is not a spectrahedral shadow for $m \geq 5$. This implies that its dual cone, the cone of copositive matrices, is not a spectrahedral shadow either.

REFERENCES

Bettiol, R.G., Kummer, M., and Mendes, R.A.E. (2021). Convex algebraic geometry of curvature operators. *SIAM J. Appl. Algebra Geom.*, 5(2), 200–228. doi:10.1137/20M1350777. URL https://doi.org/10.1137/20M1350777.

Bodnarčuk, V.G., Kalužnin, L.A., Kotov, V.N., and Romov, B.A. (1969). Galois theory for Post algebras. I, II. *Kibernetika (Kiev)*, (3), 1–10; ibid. 1969, no. 5, 1–9.

Fawzi, H. (2021). The set of separable states has no finite semidefinite representation except in dimension 3×2. *Comm. Math. Phys.*, 386(3), 1319–1335. doi:10.1007/s00220-021-04163-2. URL https://doi.org/10.1007/s00220-021-04163-2.

Hall jun., M. and Newman, M. (1963). Copositive and completely positive quadratic forms. *Proc. Camb. Philos. Soc.*, 59, 329–339.

Helton, J.W. and Nie, J. (2009). Sufficient and necessary conditions for semidefinite representability of convex hulls and sets. *SIAM J. Optim.*, 20(2), 759–791. doi:10.1137/07070526X. URL https://doi.org/10.1137/07070526X.

Nemirovski, A. (2007). Advances in convex optimization: conic programming. In *International Congress of Mathematicians. Vol. I*, 413–444. Eur. Math. Soc., Zürich. doi:10.4171/022-1/17. URL https://doi.org/10.4171/022-1/17.

Scheiderer, C. (2018). Spectrahedral shadows. *SIAM J. Appl. Algebra Geom.*, 2(1), 26–44. doi:10.1137/17M1118981.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Polarization of Multi-Agent Gradient Flows over Hypersurfaces ⋆

La Mi * **Jorge Gonçalves** *,** **Johan Markdahl** *

* *Luxembourg Centre for Systems Biomedicine*
University of Luxembourg
la.mi@uni.lu, jorge.goncalves@uni.lu, markdahl@kth.se
** *Department of Plant Sciences, University of Cambridge*

Abstract: Multi-agent systems are known to exhibit stable emergent behaviors, including polarization, over $\mathbb{R}^n$ or highly symmetric nonlinear spaces. In this article, we eschew linearity and symmetry of the underlying spaces, and study the stability of polarized equilibria of multi-agent gradient flows evolving on general hypersurfaces. The agents attract or repel each other according to the partition of the communication graph that is connected but otherwise arbitrary. The hypersurfaces are outfitted with geometric features styled "dimples" and "pimples" that characterize the absence of flatness. The signs of inter-agent couplings together with these geometric features give rise to stable polarization under various sufficient conditions.

Keywords: polarization, manifolds, multi-agent systems, gradient flows, nonlinear systems

1. INTRODUCTION

We study polarization in multi-agent gradient flow systems confined to manifolds embedded in the Euclidean space. Polarization refers to the emergent state where the agents converge to two distinct clusters. The gradient descent flow, being a sufficiently simple optimizing process, is amenable to rigorous stability analysis and thus widely adopted by many agent-based models as coordinating protocols for robot swarms. Works on nonlinear spaces is less common than those in $\mathbb{R}^n$, and are predominantly focused on highly symmetric spaces, see Sarlette and Sepulchre (2009); Sepulchre (2011); Lageman and Sun (2016). Possible applications in clustering algorithms, cellular division, and social dynamics call for studies of multi-agent systems on more general manifolds.

There are a few polarization studies on manifolds in the literature, where the n-sphere has received the most attention. Gaitonde et al. (2021) proved almost sure convergence for a class of Markov processes on the hypersphere. Hong and Strogatz (2011) found traveling wave polarization in a variant of the Kuramoto model over the unit circle with conformist and contrarian oscillators. Ha et al. (2020) derived stability conditions for a higher dimensional Kuramoto model featuring positive and negative couplings between agents. More elaborate state-dependent interaction rules inspired by neuroscience are considered by Crnkić and Jaćimović (2018) over a 3-sphere through a quaternion formulation. Another neuronal model called the principal component analysis flow is studied by Zhang et al. (2021) to obtain stability of the antipodal equilibrium. For ring graphs over the 2-sphere, Song et al. (2017) obtained asymptotically stable polarization with even number of agents. For more general manifolds, a recent work by Aydogdu et al. (2017) explored geodesic and chordal interactions between agents on general Riemannian manifolds without stability analysis.

In view of these related works, our contribution is that we provide rigorous stability analysis of polarized equilibria for the multi-agent gradient descent system with arbitrary connected network topology over more general manifolds. Our investigation is realized by outfitting the manifolds with special geometric features styled "dimples" and "pimples". The interplay between these geometric features and the cooperative/antagonistic interactions among agents then gives rise to different routes to polarization.

2. SETUP

2.1 Geometric features of the manifold

Consider a closed, connected, and orientable hypersurface embedded in the Euclidean ambient space

$$\mathcal{H}^n = \{y \in \mathbb{R}^{n+1} \,|\, c(y) = 0\}$$

implicitly characterized by a smooth C^2 function $c : \mathbb{R}^{n+1} \to \mathbb{R}$. The hypersurface $\mathcal{H}^n$ separates its complement $\mathbb{R}^{n+1} - \mathcal{H}^n$ into two disjoint sets, one where c is positive and the other where c is negative (Lima (1988)). Without loss of generality, we identify the former with the unbounded set outside $\mathcal{H}^n$, and the latter with the bounded set inside $\mathcal{H}^n$. The unit normal $n(x) = \nabla c(x)/\|\nabla c(x)\|$ is outward-pointing (i.e., pointing towards the unbounded set), where the usual restriction applies: the gradient in $\mathbb{R}^{n+1}$ satisfies $\nabla c(x) \neq 0$ for every $x \in \mathcal{H}^n$.

The hypersurface $\mathcal{H}^n$ is equipped with special features: dimples and pimples. To define them, introduce a height function $h_x \colon \mathcal{H}^n \to \mathbb{R}$ with respect to a fixed x

$$h_x(y) := \langle n(x), y \rangle, \quad \forall y \in \mathcal{H}^n.$$

The height function gives the altitude of a point y along the axis spanned by $n(x)$. For notational convenience, if

⋆ This work is funded by the Luxembourgish state agency FNR through their funding instrument CORE OPEN.

the fixed x carries a subscript, e.g. x_i, then h_{x_i} is shortened as h_i. Now we are ready to introduce the definitions for a dimple and a pimple.

Definition 1. If for some $x \in \mathcal{H}^n$, $y = x$ is a strict local minimizer of $h_x(y)$ in a sufficiently small neighborhood $\mathcal{I}_x = \{y \in \mathcal{H}^n \,|\, \|y - x\| < \epsilon\}$, then $\mathcal{I}_x$ is referred to as a dimple, and x the bottom of the dimple. Similarly, if for some $x \in \mathcal{H}^n$, $y = x$ is a strict local maximizer of $h_x(y)$ in a sufficiently small neighborhood $\mathcal{I}_x$, then $\mathcal{I}_x$ is referred to as a pimple, and x the bottom of the pimple.

Figure 1 illustrates the concepts of dimples and pimples in $\mathbb{R}^3$ by some fruits and donuts.

Remark 2.1. The features $\mathcal{I}_x$ does not necessarily contain only one bottom x. That is, another point $y \neq x$ may also be a bottom with respect to the height function h_y. It may be that $\mathcal{I}_x$ has a single set of bottoms covering all or part of $\mathcal{I}_x$, or there are multiple disjoint sets of bottoms within $\mathcal{I}_x$. We require $\mathcal{I}_x$ to be sufficiently small in Definition 1 to exclude the latter case, by shrinking the radius ϵ of the neighborhood around x. However, it is not possible to completely avoid disjoint sets of bottoms when, e.g., the embedding of the hypersurface is not analytic.

2.2 Multi-agent networks

Evolving on the hypersurface is a homogeneous multi-agent system with N agents, associated with an undirected, connected, and weighted graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$. The adjacency matrix $A = [a_{ij}]$ is symmetrical and has non-negative entries. The vertices $\mathcal{V}$ are divided into two groups $\mathcal{V}_\mathrm{u} = \{1, 2, \dots M\}$ and $\mathcal{V}_\mathrm{l} = \{M+1, \dots N\}$ for $1 < M < N$. The edge set $\mathcal{E}$ is partitioned into intragroup and intergroup sets $\mathcal{E}_+ = \{\{i, j\} \in \mathcal{E} \,|\, i, j \in \mathcal{V}_\mathrm{u} \text{ or } i, j \in \mathcal{V}_\mathrm{l}\}$ and $\mathcal{E}_- = \{\{i, j\} \in \mathcal{E} \,|\, i \in \mathcal{V}_\mathrm{l}, j \in \mathcal{V}_\mathrm{u}\}$.

Such a partition is introduced to enforce different coupling rules over edges in $\mathcal{E}_+$ and $\mathcal{E}_-$. The couplings are positive over all edges in $\mathcal{E}_+$, whereas those over $\mathcal{E}_-$ can be either all positive or all negative.

2.3 Gradient flow dynamics

Denote the states of the agents individually by x_i and collectively by $\chi := (x_i)_{i=1}^N$. The agents evolve according to a simple rule of gradient descent flow in continuous time. Given a disagreement function $V : \mathcal{H}^n \to \mathbb{R}$ with a smooth extension $\bar{V} : \mathbb{R}^{n+1} \to \mathbb{R}$, the dynamics of each agent is

$$\dot{x}_i = -\operatorname{grad}_i V(\chi) = -\mathrm{P}_i \left(\nabla_i \bar{V}(\chi)\right), \tag{1}$$

where grad_i is the intrinsic gradient on the tangent space $\mathsf{T}_{x_i}\mathcal{H}^n$ at the point x_i, $\mathrm{P}_i = I - n(x_i)n(x_i)'$ is an orthogonal projection matrix on $\mathsf{T}_{x_i}\mathcal{H}^n$, and $\nabla_i \bar{V}(\chi)$ is the standard gradient in the Euclidean space of the disagreement function.

As mentioned in §2.2, the partition of the N agents into $\mathcal{V}_\mathrm{u}$ and $\mathcal{V}_\mathrm{l}$ allows us to assign attractive or repulsive intergroup interactions. For the case of attractive intragroup coupling and repulsive intergroup coupling, the disagreement function is

$$V_-(\chi) := \frac{1}{2} \sum_{\{i,j\} \in \mathcal{E}_+} a_{ij} \|x_j - x_i\|^2 - \frac{1}{2} \sum_{\{i,j\} \in \mathcal{E}_-} a_{ij} \|x_j - x_i\|^2. \tag{2}$$

For purely attractive coupling, we simply change the sign of the second term in (2):

$$V_+(\chi) = \frac{1}{2} \sum_{\{i,j\} \in \mathcal{E}} a_{ij} \|x_j - x_i\|^2. \tag{3}$$

Substituting (2) and (3) into (1) respectively yields

$$\begin{aligned} \dot{x}_i &= \mathrm{P}_i \left(\sum_{j \in \mathcal{V}_\mathrm{u}} a_{ij} \left(x_j - x_i\right) - \sum_{j \in \mathcal{V}_\mathrm{l}} a_{ij} \left(x_j - x_i\right) \right), \quad i \in \mathcal{V}_\mathrm{u} \\ \dot{x}_i &= \mathrm{P}_i \left(\sum_{j \in \mathcal{V}_\mathrm{l}} a_{ij} \left(x_j - x_i\right) - \sum_{j \in \mathcal{V}_\mathrm{u}} a_{ij} \left(x_j - x_i\right) \right), \quad i \in \mathcal{V}_\mathrm{l} \end{aligned} \tag{4}$$

and

$$\dot{x}_i = \mathrm{P}_i \sum_{j \in \mathcal{V}} a_{ij} \left(x_j - x_i\right), \quad i \in \mathcal{V}. \tag{5}$$

2.4 The assemblage

Assembling the aforementioned ingredients in §2.2 and §2.3, we have a multi-agent gradient flow system with attractive (4) or repulsive (5) intergroup interactions evolving on a hypersurface $\mathcal{H}^n$. The hypersurface is equipped with a pair of dimples or pimples as illustrated in Fig. 1, each containing one of the two groups of agents $\mathcal{V}_\mathrm{u}$ and $\mathcal{V}_\mathrm{l}$. We are interested in possible polarization arising in this setting as a result of the interplay between the graph couplings and the geometry of the underlying nonlinear space.

Definition 2. (Polarization). The agents are said to be polarized if $x_i = x_j$ for all $\{i, j\} \in \mathcal{E}_+$ and $x_i \neq x_j$ for all $\{i, j\} \in \mathcal{E}_-$.

Definition 2 characterizes a polarized configuration without specifying whether the states are in equilibrium, limit-cycle, or other non-stationary modes. We focus on polarized equilibria and its stability properties, because gradient flows must converge to either an equilibrium or a set of equilibria (Helmke and Moore, 1996, App. C.12).

For the definitions of Lyapunov and asymptotic stability, while those of an equilibrium point are well known, those of a set of equilibria is perhaps less standard. Introduce the *Hausdorff distance* between two sets $\mathcal{Y}, \mathcal{Z} \subset \mathbb{R}^n$,

$$d_\mathrm{H}(\mathcal{Y}, \mathcal{Z}) := \max\{\sup_{y \in \mathcal{Y}} \inf_{z \in \mathcal{Z}} \|y - z\|, \sup_{z \in \mathcal{Z}} \inf_{y \in \mathcal{Y}} \|y - z\|\}.$$

Definition 3. (Stability). A set of equilibria $\mathcal{S}$ is Lyapunov stable if, for each $\epsilon > 0$, there is $\delta = \delta(\epsilon)$ such that $d_\mathrm{H}(x, \mathcal{S})|_{t=0} < \delta$ implies $d_\mathrm{H}(x, \mathcal{S})|_t < \epsilon$ for all $t \geq 0$; is asymptotically stable if it is stable and δ can be chosen such that $d_\mathrm{H}(x, \mathcal{S})|_{t=0} < \delta$ implies $\lim_{t \to \infty} d_\mathrm{H}(x, \mathcal{S}) = 0$.

We collect a few previous results and associated definitions that will pave the way for later development.

Definition 4. (Local minimizer). A set $\mathcal{S} \subset \mathcal{M}$ is said to be a *local minimizer* of a real function $f : \mathcal{M} \to \mathbb{R}$ from a metric space $(\mathcal{M}, d_\mathrm{H})$ if for some $\epsilon > 0$, there is an open neighborhood $\mathcal{N}(\mathcal{S}) = \{x \in \mathcal{M} \,|\, d_\mathrm{H}(x, \mathcal{S}) < \epsilon\}$ such that $f|_\mathcal{S} \leq f(x)$ for all $x \in \mathcal{N}(\mathcal{S})$. Moreover, if the inequality is strict for all $x \in \mathcal{N}(\mathcal{S}) \backslash \mathcal{S}$, then $\mathcal{S}$ is said to be a *strict local minimizer*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Fig. 1. Fruits and donuts illustrating Definition 1. From left to right: a lemon with a pair of pimples, an apple with a pair of dimples, a lemon donut with a pair of dimples, and an apple donut with a pair of pimples.

Definition 5. (Isolated critical). A set $\mathcal{S} \subset \mathcal{M}$ of critical points of a real function $f : \mathcal{M} \to \mathbb{R}$ from a metric space $(\mathcal{M}, d_{\mathrm{H}})$ is said to be *isolated critical* if for some $\epsilon > 0$, there is an open neighborhood $\mathcal{N}(\mathcal{S}) = \{x \in \mathcal{M} \,|\, d_{\mathrm{H}}(x, \mathcal{S}) < \epsilon\}$ such that $\mathcal{N}(\mathcal{S}) \backslash \mathcal{S}$ is void of critical points.

Proposition 6. (Prop. 3 Markdahl (2021b)). Let $\mathcal{M}$ be a closed manifold and take any $V : \mathcal{M} \to \mathbb{R}$ that is C^2. Let $\mathcal{S}$ be a compact set of local minimizers of V. If $\mathcal{S}$ is a strict local minimizer, then $\mathcal{S}$ is a Lyapunov stable equilibrium set of $\dot{x} = -\operatorname{grad} V$. If $\mathcal{S}$ is also isolated critical, then it is asymptotically stable.

3. MAIN RESULTS

In this section, we present and discuss our main results concerning the stability properties of polarized equilibria. They arise in different combinations of attractive/repulsive interactions with dimple/pimple geometric features, best exemplified in Fig. 1. Despite the symmetrical shapes exhibited in the figure, we emphasize that our general results do not require any spatial symmetry of the hypersurface embedding.

3.1 Dimple pairs with attractive intergroup couplings

Consider the setting of a pair of dimples on the hypersurface, one containing the group $\mathcal{V}_{\mathrm{u}}$ and the other $\mathcal{V}_{\mathrm{l}}$. Thus, we operate under the following assumption in this section:

Assumption 1. The sets $\mathcal{I}_{\mathrm{u}}$ and $\mathcal{I}_{\mathrm{l}}$ are a pair of dimples, and $x_i \in \mathcal{I}_{\mathrm{u}}$ for all $i \in \mathcal{V}_{\mathrm{u}}$, $x_i \in \mathcal{I}_{\mathrm{l}}$ for all $i \in \mathcal{V}_{\mathrm{l}}$.

Now we investigate stability of a polarized equilibrium that exists in the system (5) if the manifold resembles the apple in Fig. 1. The set of interest is the following:

$$\chi^* := \{\chi \in (\mathcal{H}^n)^N \,|\, x_i = x_{\mathrm{u}}\,, i \in \mathcal{V}_{\mathrm{u}}, x_i = x_{\mathrm{l}}\,, i \in \mathcal{V}_{\mathrm{l}}\}. \quad (6)$$

Let $r_o = \frac{1}{2}\|x_{\mathrm{u}} - x_{\mathrm{l}}\|$ denote the half distance between the dimple bottoms x_{u} and x_{l}.

Proposition 7. For system (5) under Assumption 1, if there exists a pair of distinct dimple bottoms $x_{\mathrm{u}} \in \mathcal{I}_{\mathrm{u}}$ and $x_{\mathrm{l}} \in \mathcal{I}_{\mathrm{l}}$ such that

(1) $x_{\mathrm{u}} - x_{\mathrm{l}}$ is parallel to $n(x_{\mathrm{u}})$, and
(2) $h_{\mathrm{u}}(x_{\mathrm{l}})$ is a local maximum satisfying $h_{\mathrm{u}}(x_{\mathrm{l}}) < h_{\mathrm{u}}(x_{\mathrm{u}})$,

then a strict local minimum of V_+ is $V_+^* := 2r_o^2 \sum_{\{i,j\} \in \mathcal{E}_-} a_{ij}$, and the corresponding strict local minimizer is χ^* defined in (6).

To prove it, we look at the disagreement contributions from $\mathcal{E}_-$ and $\mathcal{E}_+$ in (3) separately. Those from $\mathcal{E}_-$ obeys

$$\begin{aligned}\|x_j - x_i\|^2 &\geq \langle x_j - x_i, n(x_{\mathrm{u}})\rangle^2 = \big(h_{\mathrm{u}}(x_j) - h_{\mathrm{u}}(x_i)\big)^2 \\ &\geq (h_{\mathrm{u}}(x_{\mathrm{u}}) - h_{\mathrm{u}}(x_{\mathrm{l}}))^2 = 4r_o^2,\end{aligned}$$

whereas those from $\mathcal{E}_+$ are lower bounded by 0 since agents from the same group can simply converge to a single point. Both lower bounds are achieved simultaneously only by the configuration χ^* under the conditions of 7, thus the conclusion.

Lyapunov stability of (6) is then obtained by applying Prop. 6 to Prop. 7. For asymptotic stability, we need $\mathcal{I}_{\mathrm{u}}$ and $\mathcal{I}_{\mathrm{l}}$ to live on nice manifolds.

Theorem 8. For system (5) under Assumption 1, if the two dimples $\mathcal{I}_{\mathrm{u}}$ and $\mathcal{I}_{\mathrm{l}}$ satisfy the properties given in Prop. 7, and in addition, there is a neighborhood $\mathcal{N}_{\mathrm{a}}(\chi^*)$ on $(\mathcal{H}^n)^N$ that belongs to an analytic manifold, then χ^* defined in (6) is an asymptotically stable polarized equilibrium.

Proof. Following a variant (Kurdyka et al., 2000, Sec. 9) of the Łojasiewicz inequality valid on analytic Riemannian manifolds, the analytic function $V_+(\chi)$ in (3) behaves in the following way in a neighborhood of the polarized equilibrium $\mathcal{N}_{\mathrm{l}}(\chi^*) \subset \mathcal{N}_{\mathrm{a}}(\chi^*)$:

$$|V_+(\chi) - V_+(\chi^*)|^\alpha \leq \kappa \|\operatorname{grad} V_+(\chi)\|,$$

for $\alpha < 1$ and $\kappa > 0$, and where grad is the intrinsic gradient on the tangent space. Suppose that $\chi \notin \chi^*$ is an equilibrium in $\mathcal{N}_{\mathrm{l}}(\chi^*)$, then $\operatorname{grad} V_+(\chi) = 0$, c.f. (1). Consequently, $V_+(\chi) = V_+(\chi^*)$. However, Prop. 7 says that the local minimum V_+^* is achieved only by $\chi = \chi^*$, a contradiction. Therefore, there is no equilibrium except χ^* in $\mathcal{N}_{\mathrm{l}}(\chi^*)$, rendering χ^* isolated critical as per Definition 5. Thus, χ^* is asymptotically stable by Prop. 6. □

3.2 Pimple pairs with repulsive intergroup couplings

Consider a pair of pimples on the hypersurface, one containing the group of agents $\mathcal{V}_{\mathrm{u}}$ and the other $\mathcal{V}_{\mathrm{l}}$. The assumption in this section is then

Assumption 2. The sets $\mathcal{I}_{\mathrm{u}}$ and $\mathcal{I}_{\mathrm{l}}$ are a pair of pimples, and $x_i \in \mathcal{I}_{\mathrm{u}}$ for all $i \in \mathcal{V}_{\mathrm{u}}$, $x_i \in \mathcal{I}_{\mathrm{l}}$ for all $i \in \mathcal{V}_{\mathrm{l}}$.

For the dynamics, we are interested in (4) with attractive intragroup coupling and repulsive intergroup coupling corresponding to the disagreement function (2). Thus, we may picture the system (4) evolving on a lemon-like manifold in Fig. 1 left.

Denote a *closed* ball centered at a point x with radius r as $\mathcal{B}_r(x)$. Let $x_o = \frac{1}{2}(x_{\mathrm{u}} + x_{\mathrm{l}})$ denote the midpoint between x_{u}

and x_l, and recall r_o as before is the half distance between them. The following results concern the set

$$\mathcal{C}_\mathrm{lem} \coloneqq \{\chi \in (\mathcal{H}^n)^N \mid x_i = x \ \forall i \in \mathcal{V}_\mathrm{u}, x_i = y \ \forall i \in \mathcal{V}_\mathrm{l}, (x,y) \in Y\}, \tag{7}$$

where $Y \coloneqq \{(x,y) \in \mathcal{I}_\mathrm{u} \times \mathcal{I}_\mathrm{l} \,|\, \|x-y\| = 2r_o\}$. This set has at least one element $\chi^* \in \mathcal{C}_\mathrm{lem}$. It may contain other elements when, for instance, the pair of pimples belongs to a sphere.

Proposition 9. For system (4) under Assumption 2, if there exists a pair of distinct pimple bottoms $x_\mathrm{u} \in \mathcal{I}_\mathrm{u}$ and $x_\mathrm{l} \in \mathcal{I}_\mathrm{l}$ such that $\mathcal{I}_\mathrm{u}$ and $\mathcal{I}_\mathrm{l}$ are entirely contained in $\mathcal{B}_{r_o}(x_o)$, then a strict local minimum of V_- is $V_-^* \coloneqq -2r_o^2 \sum_{\{i,j\}\in\mathcal{E}_-} a_{ij}$, and the corresponding strict local minimizer is a compact set of polarized configurations $\mathcal{C}_\mathrm{lem}$ defined in (7).

Similar to the proof of Prop. 7, we look at disagreement contributions from $\mathcal{E}_-$ and $\mathcal{E}_+$ separately. The main difference is that under the condition in Prop. 9, we have $\|x_j - x_i\| \le 2r_o = \|x_\mathrm{u} - x_\mathrm{l}\|$ for all $\{i,j\} \in \mathcal{E}_-$. Likewise, Lyapunov stability can be concluded for $\mathcal{C}_\mathrm{lem}$ by applying Prop. 6 to Prop. 9.

Theorem 10. For system (4) under Assumption 2, if the two pimples $\mathcal{I}_\mathrm{u}$ and $\mathcal{I}_\mathrm{l}$ satisfy the conditions given in Prop. 9, and in addition, there is a neighborhood $\mathcal{N}_\mathrm{a}(\mathcal{C}_\mathrm{lem})$ on $(\mathcal{H}^n)^N$ that belongs to an analytic manifold, then $\mathcal{C}_\mathrm{lem}$ defined in (7) is an asymptotically stable set of polarized equilibria.

Theorem 10 concerns the asymptotic stability of an equilibrium set, rather than an equilibrium point as the singleton set χ^* in Theorem 8. The accompanying subtlety requires a more involved proof; the line of reasoning is similar to the proof of (Markdahl, 2021a, Thm. 8).

Remark 3.1. The additional requirement on the analyticity of the manifold in Theorems 8 and 10 is a local one. In fact, the sole purpose of introducing the neighborhood $\mathcal{N}_\mathrm{a}(\bullet)$ is to emphasize this local nature. We do not require the whole manifold to be analytic for $\mathcal{C}_\mathrm{lem}$ or χ^* to be asymptotically stable. For instance, $\mathcal{N}_\mathrm{a}(\mathcal{C}_\mathrm{lem})$ may be a subset of $(\mathcal{H}^n)^N \cap \mathcal{M}^N$, where $\mathcal{H}^n$ is the hypersurface on which the agents inhabit, whereas $\mathcal{M}$ is an analytic manifold.

3.3 Dimples and pimples on a torus

Stable polarization with a pair of dimples is found not only in (5) as discussed in §3.1. It also exists in the system (4) with repulsive intergroup coupling, if the normals of the two dimples point toward each other, see the lemon donut in Fig. 1. The stability conditions in the following result is identical to that in Prop. 9 and Theorem 10.

Proposition 11. For system (4) under Assumption 1, if there exists a pair of distinct dimple bottoms $x_\mathrm{u} \in \mathcal{I}_\mathrm{u}$ and $x_\mathrm{l} \in \mathcal{I}_\mathrm{l}$ such that $\mathcal{I}_\mathrm{u}$ and $\mathcal{I}_\mathrm{l}$ are entirely contained in $\mathcal{B}_{r_o}(x_o)$, then $\mathcal{C}_\mathrm{lem}$ defined in (7) is Lyapunov stable. Furthermore, if there is a neighborhood $\mathcal{N}_\mathrm{a}(\mathcal{C}_\mathrm{lem})$ on $(\mathcal{H}^n)^N$ that belongs to an analytic manifold, then $\mathcal{C}_\mathrm{lem}$ is asymptotically stable.

For the apple donut in Fig. 1 with attractive intergroup coupling, we have the following result analogous to that for the apple in §3.1, although the condition here has a switched inequality (compare to Prop. 7).

Proposition 12. For system (5) under Assumption 2, if there exists a pair of distinct pimple bottoms $x_\mathrm{u} \in \mathcal{I}_\mathrm{u}$ and $x_\mathrm{l} \in \mathcal{I}_\mathrm{l}$ such that $x_\mathrm{u} - x_\mathrm{l}$ is parallel to $n(x_\mathrm{u})$, and $h_\mathrm{u}(x_\mathrm{l})$ is a local minimum satisfying $h_\mathrm{u}(x_\mathrm{l}) > h_\mathrm{u}(x_\mathrm{u})$, then χ^* defined in (6) is Lyapunov stable. Furthermore, if there is a neighborhood $\mathcal{N}_\mathrm{a}(\chi^*)$ on $(\mathcal{H}^n)^N$ that belongs to an analytic manifold, then χ^* is asymptotically stable.

REFERENCES

Aydogdu, A., Mcquade, S., and Duteil, N. (2017). Opinion dynamics on a general compact Riemannian manifold. *Networks and Heterogeneous Media*, 12(3), 489–523.

Crnkić, A. and Jaćimović, V. (2018). Swarms on the 3-sphere with adaptive synapses: Hebbian and anti-Hebbian learning rule. *Systems & Control Letters*, 122, 32–38.

Gaitonde, J., Kleinberg, J., and Tardos, E. (2021). Polarization in geometric opinion dynamics. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, 499–519. Association for Computing Machinery.

Ha, S.Y., Kim, D., Lee, J., and Noh, S.E. (2020). Emergence of bicluster aggregation for the swarm sphere model with attractive-repulsive couplings. *SIAM Journal on Applied Dynamical Systems*, 19(2), 1225–1270.

Helmke, U. and Moore, J.B. (1996). *Optimization and dynamical systems*. Springer.

Hong, H. and Strogatz, S.H. (2011). Kuramoto model of coupled oscillators with positive and negative coupling parameters: An example of conformist and contrarian oscillators. *Physical Review Letters*, 106(5), 054102.

Kurdyka, K., Mostowski, T., and Parusinski, A. (2000). Proof of the gradient conjecture of R. Thom. *Annals of Mathematics*, 763–792.

Lageman, C. and Sun, Z. (2016). Consensus on spheres: Convergence analysis and perturbation theory. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 19–24. IEEE.

Lima, E.L. (1988). The Jordan-Brouwer separation theorem for smooth hypersurfaces. *The American Mathematical Monthly*, 95(1), 39–42.

Markdahl, J. (2021a). Counterexamples in synchronization: pathologies of consensus seeking gradient descent flows on surfaces. *Automatica*, 134, 109945.

Markdahl, J. (2021b). Synchronization on Riemannian manifolds: Multiply connected implies multistable. *IEEE Transactions on Automatic Control*, 66(9), 4311–4318.

Sarlette, A. and Sepulchre, R. (2009). Consensus optimization on manifolds. *SIAM journal on Control and Optimization*, 48(1), 56–76.

Sepulchre, R. (2011). Consensus on nonlinear spaces. *Annual reviews in control*, 35(1), 56–64.

Song, W., Markdahl, J., Zhang, S., Hu, X., and Hong, Y. (2017). Intrinsic reduced attitude formation with ring inter-agent graph. *Automatica*, 85, 193–201.

Zhang, Z., Al-Abri, S., and Zhang, F. (2021). Dissensus algorithms for opinion dynamics on the sphere. In *60th IEEE Conference on Decision and Control*, 5988–5993.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Why does strict dissipativity help in model predictive control?

Extended Abstract

Lars Grüne *

* *Mathematical Institute, University of Bayreuth, Bayreuth, Germany,*
e-mail: lars.gruene@uni-bayreuth.de

Abstract: During the last couple of years the theory of why and when Model Predictive Control (MPC) generates stable, feasible and near optimal closed-loop solutions has significantly matured. In this talk we give a survey about the contribution of the dissipativity concept in this line of research.

Keywords: Model Predictive Control, strict dissipativity, stability, near optimality, feasibility, detectability

1. INTRODUCTION

During the last couple of years the theory of why and when Model Predictive Control (MPC) generates stable, feasible and near optimal closed-loop solutions has significantly matured. In this talk we give a survey about the contribution of the dissipativity concept in this line of research.

2. PROBLEM FORMULATION

We present our results for discrete-time nonlinear control systems of the form

$$x(k+1) = f(x(k), u(k)), \quad x(0) = x_0 \tag{1}$$

with $x(k) \in X$ and $u(k) \in U$ for normed vector spaces X and U. Most of the results in this talk hold in an analogous way for continuous time systems.

MPC then computes a control input $u_{MPC}(\cdot)$ by solving a sequence of optimal control problems on finite, overlapping time horizons. Here, the finite horizon optimal control problem is given as follows. For a given constraint set $\mathbb{Y} \subset X \times U$, a terminal constraints set $\mathbb{X}_f$, a stage cost $\ell : \mathbb{Y} \to \mathbb{R}$, a terminal cost $F : \mathbb{X}_f \to \mathbb{R}$, and a time horizon $N \in \mathbb{N}$ we define the finite horizon functional

$$J_N(x_0, u(\cdot)) := \sum_{k=0}^{N-1} \ell(x(k), u(k)) + F(x(N)), \tag{2}$$

where $x(\cdot)$ solves (1). Then we solve

$$\text{minimize}_{u(\cdot)} J_N(x_0, u(\cdot)) \tag{3}$$

subject to the constraints $(x(k), u(k)) \in \mathbb{Y}$ for all $k = 0, \ldots, N-1$ and $x(N) \in \mathbb{X}_f$. We call a control $u(\cdot)$ *admissible* (for x_0) when these constraints are satisfied. Moreover, we set $\mathbb{X} := \{x \in X \,|\, \text{there is } u \in U \text{ with } (x,u) \in \mathbb{Y}\}$.

The pair $(\mathbb{X}_f, F)$ is referred to as *terminal condition* and in the trivial case $\mathbb{X}_f = \mathbb{X}$ and $F \equiv 0$ we refer to (2) as a problem *without terminal conditions*.

Associated to the optimal control problems (3) we define the *optimal value function*

$$V_N(x_0) := \inf_{u(\cdot) \text{ admissible}} J_N(x_0, u(\cdot))$$

and we call an admissible control $u^*(\cdot)$ *optimal* (for x_0), if $J_N(x_0, u^*(\cdot)) = V_N(x_0)$.

The corresponding MPC scheme then reads as follows (for much more detailed expositions we refer to Rawlings et al. (2017); Grüne and Pannek (2017)):

Given an initial condition $x_{MPC}(0) := \hat{x}_0 \in \mathbb{X}$ and an optimisation horizon $N \in \mathbb{N}$, for $n = 0, 1, 2, \ldots$ we perform the following steps:

(1) Let $x_0 := x_{MPC}(n)$ denote the current state of the system.
(2) Solve the finite horizon optimal control problem (3) in order to obtain the optimal control sequence $u^*(\cdot)$.
(3) Apply the first element of the optimal control sequence $u^*(\cdot)$ as a feedback control value until the next time instant, i.e., set $u_{MPC}(n) := u^*(0)$ and $x_{MPC}(n+1) := f(x_{MPC}(n), u^*(0))$.
(4) Set $n := n+1$ and go to Step 1.

Here, we consider general cost functions ℓ that do not need to have any a priori structure. This setting is typically termed *economic* MPC in the literature, although *general* MPC might be a more appropriate name.

When dealing with MPC, some of the central questions are:

- **Stability:** Does the MPC closed-loop solution exhibit stable behaviour?

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

- **Optimality:** Does the MPC closed-loop solution enjoy (approximate) optimality properties?

- **Feasibility:** Does the MPC closed-loop solution maintain the constraints and are the optimal control problems in Step (2) of the algorithm always feasible?

As we will explain in the next section, a suitable dissipativity concept helps to give positive answers to all questions.

3. STRICT DISSIPATIVITY

The appropriate dissipativity concept is the following strict dissipativity notion. In this extended abstract we limit ourselves to strict dissipativity at an equilibrium $(x^e, u^e) \in \mathbb{Y}$ (i.e., $f(x^e, u^e) = x^e$). Extensions to periodic and general time-varying trajectories are possible and will be briefly explained in the talk.

Definition 3.1. The optimal control problem is called *strictly dissipative* at an equilibrium (x^e, u^e), if there exists a *storage function* $\lambda : \mathbb{X} \to \mathbb{R}$, bounded from below, and a function [1] $\alpha \in \mathcal{K}_\infty$ such that the inequality

$$\lambda(f(x,u)) \le \lambda(x) + \ell(x,u) - \ell(x^e,u^e) - \alpha(\|x - x^e\|)$$

holds for all $(x,u) \in \mathbb{Y}$ with $f(x,u) \in \mathbb{X}$. Here, the function $s(x,u) = \ell(x,u) - \ell(x^e,u^e)$ is called the supply rate.

The optimal control problem is called *dissipative* if the above inequality holds with $\alpha \equiv 0$.

It follows immediately that if (not necessarily strict) dissipativity holds, then (x^e, u^e) is an optimal equilibrium, in the sense that $\ell(x^e,u^e) \le \ell(\tilde{x},\tilde{u})$ for all equilibria $(\tilde{x},\tilde{u}) \in \mathbb{Y}$ with $\tilde{x} \ne x^e$.

The dissipativity notion for control systems was introduced by Willems (1972) in continuous time, the discrete time version used here is due to Byrnes and Lin (1994). It is interesting that *strict* dissipativity has not played a significant role in the literature until quite recently. The reason is that in the past the specific form of the suppy function often did not play a role. In this case, any dissipative system is also strictly dissipative; it suffices to replace $s(x,u)$ by $s(x,u) + \alpha(\|x - x^e\|)$. However, if the supply function s is linked to the cost function of the optimal control problem as in Definition 1, then it is not possible to modify it. In this sense, the application to MPC and optimal control are probably the main motivation for studying *strict* dissipativity.

4. STABILITY AND AVERAGED OPTIMALITY

The observation that dissipativity is beneficial for MPC was first made in Diehl et al. (2011), where it was observed that strict duality — which is nothing but strict dissipativity with a linear storage function — implies asymptotic stability of the optimal equilibrium for the MPC closed-loop under appropriate terminal conditions. This paper already contains the key idea of all dissipativity-based MPC stability results, namely the fact that the optimal value function of the optimal control problem with rotated cost

$$\tilde{\ell}(x,u) = \ell(x,u) - \ell(x^e,u^e) + \lambda(x) - \lambda(f(x,u))$$

can be used as a Lyapunov function for the closed loop. The observation that this construction can be extended without additional effort from strict duality to strict dissipativity was then made in Angeli and Rawlings (2010).

The decisive contribution of the terminal condition in these papers lies in the fact that under this condition the optimal trajectories of the optimal control problems with cost ℓ and $\tilde{\ell}$, respectively, coincide. The properties of the terminal conditions needed for this were given in Amrit et al. (2011) and a special case was already used earlier in Angeli et al. (2009) in order to prove average optimality of the MPC closed-loop, i.e., that

$$\overline{J}_\infty(\hat{x}_0, u_{MPC}(\cdot)) = \inf_{u \text{ admissible}} \overline{J}_\infty(\hat{x}_0, u).$$

for $\overline{J}_\infty(x_0,u) := \lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \ell(x(k),u(k))$. We remark that, in contrast to most other results discussed here, for this proof strict dissipativity is not needed. However, it needs optimal operation of the system at the equilibrium (x^e, u^e), which under a controllability condition implies (non-strict) dissipativity, see Müller (2014).

The fact that the optimal trajectories with cost ℓ and $\tilde{\ell}$ coincide is no longer the case for MPC without terminal conditions. However, as first observed in Grüne (2013), then refined in Grüne and Stieler (2014) and later streamlined in Chapter 8 of Grüne and Pannek (2017), the solutions are still very similar up to a certain time P. The reason for this is the so-called *turnpike property* in optimal control, which demands that the optimal trajectory most of the time stays near the optimal equilibrium. As noted in Grüne (2013), this property is implied by strict dissipativity under a reachability condition (conceptually similar results are much older and can be found, e.g., in Carlson et al. (1991)). Besides the possibility of building a Lyapunov function, its implication of the turnpike property is the second important feature of strict dissipativity. For more information on the turnpike property we refer to the recent survey Faulwasser and Grüne (2022).

As a consequence of this similarity, without terminal conditions we can still conclude near average optimality, i.e.,

$$\overline{J}_\infty(\hat{x}_0, u_{MPC}(\cdot)) = \inf_{u \text{ admissible}} \overline{J}_\infty(\hat{x}_0, u) + \varepsilon(N)$$

with $\varepsilon(N) \to 0$ as $N \to \infty$, and practical asymptotic stability of the closed loop, i.e., asymptotically stable behaviour outside a small neighbourhood of x^e, whose size also tends to 0 as N tends to infinity. This is due to the fact that the optimal value function for cost $\tilde{\ell}$ is still an approximate Lyapunov function for the MPC closed loop. These two properties hold provided the optimal value functions for different time horizons satisfy a uniform continuity condition at the optimal equilibrium x^e, which is needed in order to avoid that the small differences in the optimal trajectories cause large differences in the closed-loop behaviour.

[1] As usual, we define $\mathcal{K}_\infty$ to be the space of continuous functions $\alpha : [0,\infty) \to [0,\infty)$ with $\alpha(0) = 0$ and α is strictly increasing to ∞.

5. TRANSIENT OPTIMALITY

While average optimality is a good measure to assess the performance of trajectories on very long time horizons, it does not tell much on short horizons. The reason is that a large cost on a short horizon contributes only very little to the average over a long horizon. Hence, a low average cost on a very long horizon does not allow for any conclusions about the cost on short horizons of the same trajectory. To this end, the concept of *transient optimality* is useful. Recall that under the strict dissipativity condition the closed-loop solutions converge to x^e (with appropriate terminal conditions) or to a small neighbourhood thereof (without terminal conditions). Hence, if we fix a sufficiently large time $K \in \mathbb{N}$ (that may be much larger than N), then we can find a small $\varepsilon > 0$ such that $\|x_{MPC}(n) - x^e\| \leq \varepsilon$ for all $n \geq K$. We can now compare the cost of this trajectory up to time K, i.e.,

$$J_K(\hat{x}_0, u_{MPC}(\cdot))$$

with the cost of all other trajectories that also end up in an ε-neighbourhood of x^e, i.e., with $V_K^{tr}(x_0) :=$

$$\inf\{J_K(x_0, u(\cdot)) \,|\, u(\cdot) \text{ admissible}, \|x(K) - x^e\| \leq \varepsilon\}.$$

It turns out that there exist two functions $\varepsilon_1(N), \varepsilon_2(K) \to 0$ as $N, K \to \infty$, such that

$$J_K(\hat{x}_0, u_{MPC}(\cdot)) \leq V_K^{tr}(\hat{x}_0) + \varepsilon_1(N) + \varepsilon_2(K)$$

in the case with terminal conditions and

$$J_K(\hat{x}_0, u_{MPC}(\cdot)) \leq V_K^{tr}(\hat{x}_0) + K\varepsilon_1(N) + \varepsilon_2(K)$$

in the case without terminal conditions. The former was proved in Grüne and Panin (2015) and the latter in Grüne and Stieler (2014); a unified treatment of both cases was later given in (Grüne and Pannek, 2017, Chapter 8).

6. FEASIBILITY

In all statements so far we have tacitly assumed that the solution $x_{MPC}(n)$ exists for all $n \geq 0$. However, this requires that in each sampling instance in Step (2) of the MPC scheme there exists an admissible control $u(\cdot)$ for the initial condition $x_0 = x_{MPC}(n)$. In this case, we call $x_0 = x_{MPC}(n)$ *feasible* and the question is thus whether $x_{MPC}(n)$ is feasible for all $n \geq 0$ In case of MPC with terminal conditions, feasibility for $x_{MPC}(n)$ follows if $x_{MPC}(n-1)$ is feasible — a property called *recursive feasibility* — provided the terminal constrained $\mathbb{X}_f$ is viable, i.e., for each $x \in \mathbb{X}_f$ there is $u \in U$ with $(x, u) \in \mathbb{Y}$ and $f(x, u) \in \mathbb{X}_f$, see, e.g., Mayne et al. (2000). This procedure and the related proofs are completely unrelated to dissipativity.

However, in the absence of terminal conditions, strict dissipativity or, more precisely, the turnpike property again play an important role. If we assume that the optimal equilibrium x^e lies in the interior of the state constraint set $\mathbb{X}$, then for all sufficiently large horizons N the turnpike property implies feasibility for all points that lies on the part of the optimal trajectory that approaches the turnpike. From this observation, it is then possible to conclude recursive feasibility, see Faulwasser and Bonvin (2015); Faulwasser et al. (2018).

7. CONCLUSION AND RECENT DEVELOPMENTS

Strict dissipativity allows to conclude a variety of desirable properties for the closed-loop system generated by MPC schemes with general cost functions. The two decisive features of strict dissipativity in the context of MPC are (i) that it allows to build a Lyapunov function for the closed-loop based on an optimal control problem with cost $\tilde{\ell}$ and (ii) that it implies the turnpike property.

This has motivated extensive studies about the nature of strict dissipativity. A very interesting connection for linear quadratic problems is that strict dissipativity is closely related to detectability properties, see Grüne and Guglielmi (2018, 2021), which in turn are again closely linked to the turnpike property in a very general infinite-dimensional evolution equation setting, see Grüne et al. (2019, 2020, 2021). This relation will also be explained in the talk. Dissipativity also turned out to be very useful for understanding the long-term behavior of infinite-horizon optimal control problems, see Faulwasser and Kellett (2021).

REFERENCES

Amrit, R., Rawlings, J.B., and Angeli, D. (2011). Economic optimization using model predictive control with a terminal cost. *Annual Rev. Control*, 35, 178–186.

Angeli, D., Amrit, R., and Rawlings, J.B. (2009). Receding horizon cost optimization for overly constrained nonlinear plants. In *Proceedings of the 48th IEEE Conference on Decision and Control – CDC 2009*, 7972–7977. Shanghai, China.

Angeli, D. and Rawlings, J.B. (2010). Receding horizon cost optimization and control for nonlinear plants. In *Proceedings of the 8th IFAC Symposium on Nonlinear Control Systems – NOLCOS 2010*, 1217–1223. Bologna, Italy.

Byrnes, C.I. and Lin, W. (1994). Losslessness, feedback equivalence, and the global stabilization of discrete-time nonlinear systems. *IEEE Trans. Automat. Control*, 39(1), 83–98.

Carlson, D.A., Haurie, A.B., and Leizarowitz, A. (1991). *Infinite horizon optimal control — Deterministic and Stochastic Systems*. Springer-Verlag, Berlin, second edition.

Diehl, M., Amrit, R., and Rawlings, J.B. (2011). A Lyapunov function for economic optimizing model predictive control. *IEEE Trans. Autom. Control*, 56, 703–707.

Faulwasser, T. and Bonvin, D. (2015). On the design of economic NMPC based on approximate turnpike properties. In *Proceedings of the 54th IEEE Conference on Decision and Control — CDC 2015*, 4964–4970.

Faulwasser, T. and Grüne, L. (2022). Turnpike properties in optimal control: an overview of discrete-time and continuous-time results. To appear in the Handbook of Numerical Analysis. Paper available from https://doi.org/10.1016/bs.hna.2021.12.011. Preprint available from https://arxiv.org/pdf/2011.13670.pdf.

Faulwasser, T., Grüne, L., and Müller, M.A. (2018). Economic nonlinear model predictive control. *Foundations and Trends® in Systems and Control*, 5(1), 1–98.

Faulwasser, T. and Kellett, C.M. (2021). On continuous-time infinite horizon optimal control—dissipativity, stability, and transversality. *Automatica*, 134. doi: 10.1016/j.automatica.2021.109907. Paper No. 109907.

Grüne, L. (2013). Economic receding horizon control without terminal constraints. *Automatica*, 49(3), 725–734.

Grüne, L. and Guglielmi, R. (2018). Turnpike properties and strict dissipativity for discrete time linear quadratic optimal control problems. *SIAM J. Cont. Optim.*, 56(2), 1282–1302.

Grüne, L. and Guglielmi, R. (2021). On the relation between turnpike properties and dissipativity for continuous time linear quadratic optimal control problems. *Math. Control Rel. Fields*, 11(1), 169–188.

Grüne, L. and Panin, A. (2015). On non-averaged performance of economic MPC with terminal conditions. In *Proceedings of the 54th IEEE Conference on Decision and Control — CDC 2015*, 4332–4337. Osaka, Japan.

Grüne, L. and Pannek, J. (2017). *Nonlinear Model Predictive Control. Theory and Algorithms*. Springer-Verlag, London, 2nd edition.

Grüne, L., Schaller, M., and Schiela, A. (2019). Sensitivity analysis of optimal control for a class of parabolic PDEs motivated by model predictive control. *SIAM J. Control Optim.*, 57(4), 2753–2774.

Grüne, L., Schaller, M., and Schiela, A. (2020). Exponential sensitivity and turnpike analysis for linear quadratic optimal control of general evolution equations. *J. Differ. Equ.*, 268(12), 7311–7341.

Grüne, L., Schaller, M., and Schiela, A. (2021). Abstract nonlinear sensitivity and turnpike analysis and an application to semilinear parabolic PDEs. *ESAIM COCV*, 27. doi:10.1051/cocv/2021030. Paper No. 56, 28 pages.

Grüne, L. and Stieler, M. (2014). Asymptotic stability and transient optimality of economic MPC without terminal conditions. *J. Proc. Control*, 24(8), 1187–1196.

Mayne, D.Q., Rawlings, J.B., Rao, C.V., and Scokaert, P.O.M. (2000). Constrained model predictive control: stability and optimality. *Automatica*, 36, 789–814.

Müller, M.A. (2014). *Distributed and economic model predictive control: beyond setpoint stabilization*. Ph.D. thesis, Universität Stuttgart, Germany.

Rawlings, J.B., Mayne, D.Q., and Diehl, M.M. (2017). *Model Predictive Control: Theory, Computation and Design*. Nob Hill Publishing, Madison, Wisconsin.

Willems, J.C. (1972). Dissipative dynamical systems. I. General theory. *Arch. Rational Mech. Anal.*, 45, 321–351.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Algebraic and geometrical aspects of cyclic subspace codes ⋆

Ferdinando Zullo *

* *Università degli Studi della Campania "Luigi Vanvitelli", 81012 Caserta, Italy (e-mail: ferdinando.zullo@unicampania.it).*

Abstract: Cyclic subspace codes gained a lot of attention especially because they may be used in random network coding for correction of errors and erasures. Roth, Raviv and Tamo in 2018 established a connection between cyclic subspace codes (with certain parameters) and Sidon spaces, introduced by Bachoc, Serra and Zémor in 2017 in relation with the linear analogue of Vosper's Theorem. This connection allowed Roth, Raviv and Tamo to construct large classes of cyclic subspace codes with one or more orbits. In this abstract we will consider an extension of the notion of Sidon space, which turns out to be related to cyclic subspace codes with more than one orbit. Moreover, we will also use the geometry of linear sets to provide some bounds on the parameters of a cyclic subspace code.

Keywords: cyclic subspace code; Sidon space; multi-Sidon space; linear set; orbit.

1. INTRODUCTION

Let k be a non-negative integer with $k \leq n$, the set of all k-dimensional $\mathbb{F}_q$-subspaces of $\mathbb{F}_{q^n}$, viewed as $\mathbb{F}_q$-vector space, forms a **Grassmanian space** over $\mathbb{F}_q$, which is denoted by $\mathcal{G}_q(n,k)$. A **constant dimension subspace code** is a subset C of $\mathcal{G}_q(n,k)$ endowed with the metric defined as follows

$$d(U,V) = \dim_{\mathbb{F}_q}(U) + \dim_{\mathbb{F}_q}(V) - 2\dim_{\mathbb{F}_q}(U \cap V),$$

where $U, V \in C$. This metric is also known as **subspace metric**. Subspace codes have been recently used for the error correction in random network coding, see Koetter and Kschischang (2008). The first class of subspace codes studied was the one introduced in Etzion and Vardy (2011), which is known as **cyclic subspace codes**. A subspace code $C \subseteq \mathcal{G}_q(n,k)$ is said to be **cyclic** if for every $\alpha \in \mathbb{F}_{q^n}^*$ and every $V \in C$ then $\alpha V \in C$.

Let $V \in \mathcal{G}_q(n,k)$, the **orbit** of V is the set $C_V = \{\alpha V : \alpha \in \mathbb{F}_{q^n}^*\}$, and its size is $(q^n-1)/(q^t-1)$, for some t which divides n, see (Otal and Özbudak, 2017, Theorem 1).

In particular, every orbit of a subspace $V \in \mathcal{G}_q(n,k)$ defines a cyclic subspace code of size $(q^n-1)/(q^t-1)$, for some $t \mid n$. From now on, assume $k > 1$. Clearly, a cyclic subspace code generated by an orbit of a subspace V with size $(q^n-1)/(q-1)$ has minimum distance at most $2k-2$ and in Trautmann et al. (2013) the authors conjectured the existence of a cyclic code of size $\frac{q^n-1}{q-1}$ in $\mathcal{G}_q(n,k)$ and minimum distance $2k-2$ for every positive integers n, k such that $1 < k \leq n/2$.

⋆ The research of Ferdinando Zullo was supported by the project "VALERE: VAnviteLli pEr la RicErca" of the University of Campania "Luigi Vanvitelli" and was partially supported by the Italian National Group for Algebraic and Geometric Structures and their Applications (GNSAGA - INdAM).

In Ben-Sasson et al. (2016) the authors used subspace polynomials to generate cyclic subspace codes with size $\frac{q^n-1}{q-1}$ and minimum distance $2k-2$, proving that the conjecture is true for any given k and infinitely many values of n. Such result was then improved in Otal and Özbudak (2017). Finally, the conjecture was solved in Roth et al. (2017) for most of the cases, by making use of Sidon spaces originally introduced in Bachoc et al. (2017).

The connection between cyclic subspace codes and Sidon spaces relies on the following result.

Theorem 1. (Roth et al., 2017, Lemma 34) Let U be an $\mathbb{F}_q$-subspace of $\mathbb{F}_{q^n}$ of dimension t. Then C_U is a cyclic subspace code of size $\frac{q^n-1}{q-1}$ and minimum distance $2t-2$ if and only if U is a Sidon space.

In this abstract we provide a generalization of the notion of Sidon space via Theorem 1, introducing the multi-Sidon space notion, which is a collection of $\mathbb{F}_q$-subspaces in $\mathbb{F}_{q^n}$ with a special patterns of intersection with the elements of their orbits. We then show a link between multi-Sidon spaces of maximum dimension and cyclic subspace codes with certain parameters, which yields to a canonical form for such codes. Then we propose a geometric interpretation of the Sidon (and the multi-Sidon) property by means of linear sets, which will give us an upper bound on the number of subspaces that a cyclic subspace codes associated with a multi-Sidon space can have. The results rely on the paper Zullo (2021).

2. PRELIMINARIES

2.1 Linear sets

A point set L of $\Lambda = \mathrm{PG}(V, \mathbb{F}_{q^n}) = \mathrm{PG}(r-1, q^n)$ is said to be an $\mathbb{F}_q$**-linear set** of Λ of rank k if it is defined by the

non-zero vectors of a k-dimensional $\mathbb{F}_q$-vector subspace U of V, i.e.

$$L = L_U := \{\langle \mathbf{u} \rangle_{\mathbb{F}_{q^n}} : \mathbf{u} \in U \setminus \{\mathbf{0}\}\}.$$

We denote the rank of an $\mathbb{F}_q$-linear set L_U by $\mathrm{Rank}(L_U)$. For any subspace $S = \mathrm{PG}(Z, \mathbb{F}_{q^n})$ of Λ, the **weight** of S in L_U is defined as $w_{L_U}(S) = \dim_{\mathbb{F}_q}(U \cap Z)$. If N_i denotes the number of points of Λ having weight $i \in \{0, \dots, k\}$ in L_U, the following relations hold:

$$|L_U| \leq \frac{q^k - 1}{q - 1}, \tag{1}$$

$$|L_U| = N_1 + \ldots + N_k, \tag{2}$$

$$N_1 + N_2 \frac{q^2 - 1}{q - 1} + \ldots + N_k \frac{q^k - 1}{q - 1} = \frac{q^k - 1}{q - 1}. \tag{3}$$

For further details on linear sets see Lavrauw and Van de Voorde (2015); Polverino (2010).

2.2 Sidon spaces

Sidon spaces were introduced recently in Bachoc et al. (2017) as an important tool to prove the linear analogue of Vosper's Theorem, which analyze the equality in the linear analouge of Kneser's theorem proved in Bachoc et al. (2018); Hou et al. (2002). An $\mathbb{F}_q$-subspace U of $\mathbb{F}_{q^n}$ is called a **Sidon space** if the product of any two elements of U has a unique factorization over U, up to multiplying by some elements in $\mathbb{F}_q$. Formally, U is a Sidon space if for all nonzero $a, b, c, d \in U$, if $ab = cd$, then

$$\{a\mathbb{F}_q, b\mathbb{F}_q\} = \{c\mathbb{F}_q, d\mathbb{F}_q\},$$

where if $e \in \mathbb{F}_{q^n}$ then $e\mathbb{F}_q = \{e\lambda : \lambda \in \mathbb{F}_q\}$. Sidon spaces may be seen as the q-analogue of **Sidon sets**, see O'Bryant (2004).

3. MULTI-SIDON SPACES AND CYCLIC SUBSPACE CODES

Although Sidon spaces are defined purely algebraically, they can be defined (via Theorem 1) as those subspaces of $\mathbb{F}_{q^n}$ meeting each elements of its orbit in dimension at most one. Taking this into account, the following definition arises naturally.

Let $\{U_1, \dots, U_r\}$ be a set of $\mathbb{F}_q$-subspaces of $\mathbb{F}_{q^n}$ such that $\mathrm{Orb}(U_i) \cap \mathrm{Orb}(U_j) = \emptyset$, for every $i, j \in \{1, \dots, r\}$ with $i \neq j$. Let $k_i = \dim_{\mathbb{F}_q}(U_i) \geq 2$ for any $i \in \{1, \dots, r\}$ and suppose that $|\mathrm{Orb}(U_i)| = \frac{q^n - 1}{q - 1}$ for every $i \in \{1, \dots, r\}$. If $\dim_{\mathbb{F}_q}(U_i \cap \alpha U_j) \leq 1$, for every $\alpha \in \mathbb{F}_{q^n}$ and $i, j \in \{1, \dots, r\}$ such that $U_i \neq \alpha U_j$ then we call $\{U_1, \dots, U_r\}$ a **multi-Sidon space**.

Clearly, when $r = 1$ a multi-Sidon space is a Sidon space. Moreover, if $\{U_1, \dots, U_r\}$ is a multi-Sidon space then U_i is a Sidon space for every $i \in \{1, \dots, r\}$.

Let U and V be two $\mathbb{F}_q$-subspaces of $\mathbb{F}_{q^n}$. Denote by $\langle U^2 \rangle$ the $\mathbb{F}_q$-span of $\{st : s, t \in U\}$, $U^{-1} = \{u^{-1} : u \in U \setminus \{0\}\}$ and $U \cdot V = \{uv : u \in U, v \in V\}$.

Bachoc, Serra and Zémor proved a lower bound on the dimension of $\langle U^2 \rangle$ when U is a Sidon space in (Bachoc et al., 2017, Theorem 18) and hence, putting together with the trivial upper bound on the dimension of $\langle U^2 \rangle$, the following result holds.

Theorem 2. If U is a Sidon space in $\mathbb{F}_{q^n}$ of dimension $k \geq 3$, then

$$2k \leq \dim_{\mathbb{F}_q}(\langle U^2 \rangle) \leq \binom{k+1}{2}.$$

We can hence apply the above result to all the subspaces of a multi-Sidon space, obtaining the following bounds.

Corollary 3. Let $\{U_1, \dots, U_r\}$ is a multi-Sidon space of $\mathbb{F}_{q^n}$. Let $k_i = \dim_{\mathbb{F}_q}(U_i)$ for any $i \in \{1, \dots, r\}$. Then

$$2\sum_{i=1}^{r} k_i \leq \dim_{\mathbb{F}_q}(\langle U_1^2 \rangle \times \ldots \times \langle U_r^2 \rangle) \leq \sum_{i=1}^{r} \binom{k_i + 1}{2}.$$

In particular, $k_i \leq n/2$ for each $i \in \{1, \dots, r\}$.

Let $\{U_1, \dots, U_r\}$ be a multi-Sidon space of $\mathbb{F}_{q^n}$ and let $k_i = \dim_{\mathbb{F}_q}(U_i)$ for any $i \in \{1, \dots, r\}$. Similarly to (Roth et al., 2017, Definition 4), we say that $\{U_1, \dots, U_r\}$ is **minimum-span** if $\dim_{\mathbb{F}_q}(\langle U_1^2 \rangle \times \ldots \times \langle U_r^2 \rangle) = 2\sum_{i=1}^{r} k_i$ and we say that it is **maximum-span** if $\dim_{\mathbb{F}_q}(\langle U_1^2 \rangle \times \ldots \times \langle U_r^2 \rangle) = \sum_{i=1}^{r} \binom{k_i+1}{2}$.

A multi-Sidon space $\{U_1, \dots, U_r\}$ of $\mathbb{F}_{q^n}$, with $k_i = \dim_{\mathbb{F}_q}(U_i)$ for any $i \in \{1, \dots, r\}$, is said to be **maximum** if n is even and $k_i = n/2$ for every $i \in \{1, \dots, r\}$.

Denote by $\mathcal{L}_t$ the following set of polynomials

$$\mathcal{L}_t = \left\{ \sum_{i=0}^{t-1} a_i x^{q^i} : a_i \in \mathbb{F}_{q^t} \right\} \subset \mathbb{F}_{q^t}[x].$$

This polynomials are known as **linearized polynomials**, see Wu and Liu (2013) for more details.

We can give a canonical form for a maximum multi-Sidon space, see (Zullo, 2021, Theorem 4.5) and also (Napolitano et al., 2021, Theorem 4.6).

Theorem 4. Let $n = 2t$ and suppose that $\mathcal{U} = \{U_1, \dots, U_r\}$ is a set of $\mathbb{F}_q$-subspaces in $\mathbb{F}_{q^n}$ with dimension t. Then, up to replacing the subspaces of $\mathcal{U}$ by an element of the relative orbit, $\mathcal{U}$ coincides with

$$\{W_{f_1,\eta_1}, \dots, W_{f_r,\eta_r}\},$$

where $f_1(x), \dots, f_r(x) \in \mathcal{L}_t$, $\eta_1, \dots, \eta_r \in \mathbb{F}_{q^n} \setminus \mathbb{F}_{q^t}$ and

$$W_{f_i,\eta_i} = \{x + \eta_i f_i(x) : x \in \mathbb{F}_{q^t}\},$$

for every $i \in \{1, \dots, r\}$. Let $\eta_i = A_{i,j}\eta_j + B_{i,j}$ and $\eta_i^2 = a_i\eta_i + b_i$ with $A_{i,j}, B_{i,j}, a_i, b_i \in \mathbb{F}_{q^t}$ for any $i, j \in \{1, \dots, r\}$. Moreover, $\mathcal{U}$ is a multi-Sidon space if and only if for every $i, j \in \{1, \dots, r\}$ and $\alpha_0, \alpha_1 \in \mathbb{F}_{q^t}$ the following linearized polynomials in $\mathcal{L}_t$

$$F_{i,j}(x) = f_i(\alpha_0 x) + f_i(\alpha_1 A_{j,i} b f_j(x)) + f_i(\alpha_0 B_{j,i} f_j(x))$$
$$-\alpha_1 x - \alpha_0 A_{j,i} f_j(x) - \alpha_1 A_{j,i} a_i f_j(x) - \alpha_1 B_{j,i} f_j(x)$$

have at most q roots over $\mathbb{F}_{q^t}$.

For multiple orbits codes we have the following connection with multi-Sidon spaces, which immediately follows by the definition of the subspace metric.

Proposition 5. Let $n = 2t$ and let $U_1, \dots, U_r$ be $\mathbb{F}_q$-subspaces of dimension t in $\mathbb{F}_{q^n}$ and let

$$C = \bigcup_{i \in \{1,\dots,r\}} C_{U_i} \subseteq \mathcal{G}_q(n,t)$$

be a subspace code. Then C is a cyclic subspace code of size $r\frac{q^n-1}{q-1}$ and minimum distance $2t-2$ if and only if $\{U_1, \dots, U_r\}$ is a multi-Sidon space.

Proof. First suppose that C is a cyclic subspace code of size $r\frac{q^n-1}{q-1}$ and minimum distance $2t-2$. Observe that the size of the orbits of the U_i's is $\frac{q^n-1}{q-1}$ for every i, since the code C has size $r\frac{q^n-1}{q-1}$. The minimum distance of C is $2t-2$ which implies that

$$\dim_{\mathbb{F}_q}(U_i \cap \alpha U_j) \leq 1$$

for every $\alpha \in \mathbb{F}_{q^n}$ and $i, j \in \{1, \dots, r\}$ with $i \neq j$ and for every $\alpha \in \mathbb{F}_{q^n} \setminus \mathbb{F}_q$ if $i = j$, that is $\{U_1, \dots, U_r\}$ is a multi-Sidon space. Conversely, assume that $\{U_1, \dots, U_r\}$ is a multi-Sidon space. By construction, the code C has size $r\frac{q^n-1}{q-1}$ and its minimum distance is less than or equal to $2t-2$. The proof concludes once we note that there exists an element $\alpha \in \mathbb{F}_{q^n} \setminus \mathbb{F}_q$ such that $\dim_{\mathbb{F}_q}(U_i \cap \alpha U_j) = 1$ because of its orbit size.

Combining Theorem 4 and Proposition 5 we obtain the following canonical form for cyclic subspace code whose subspaces have dimension $n/2$, with the description of a condition equivalent to the request of having minimum distance $2k-2$.

Corollary 6. Let $n = 2t$ and let $U_1, \dots, U_r$ be $\mathbb{F}_q$-subspaces of dimension t in $\mathbb{F}_{q^n}$ and let

$$C = \bigcup_{i \in \{1,\dots,r\}} C_{U_i} \subseteq \mathcal{G}_q(n,t)$$

be a subspace code. Then

$$C = \bigcup_{i \in \{1,\dots,r\}} W_{f_i,\eta_i},$$

where $f_1(x), \dots, f_r(x) \in \mathcal{L}_t$, $\eta_1, \dots, \eta_r \in \mathbb{F}_{q^n} \setminus \mathbb{F}_{q^t}$ and

$$W_{f_i,\eta_i} = \{x + \eta_i f_i(x) \colon x \in \mathbb{F}_{q^t}\},$$

for every $i \in \{1, \dots, r\}$. Let $\eta_i = A_{i,j}\eta_j + B_{i,j}$ and $\eta_i^2 = a_i\eta_i + b_i$ with $A_{i,j}, B_{i,j}, a_i, b_i \in \mathbb{F}_{q^t}$ for any $i, j \in \{1, \dots, r\}$. Moreover, C has minimum distance $2t-2$ if and only if for every $i, j \in \{1, \dots, r\}$ and $\alpha_0, \alpha_1 \in \mathbb{F}_{q^t}$ the following linearized polynomials in $\mathcal{L}_t$

$$\begin{aligned} F_{i,j}(x) = f_i(\alpha_0 x) + f_i(\alpha_1 A_{j,i} b f_j(x)) + f_i(\alpha_0 B_{j,i} f_j(x)) \\ -\alpha_1 x - \alpha_0 A_{j,i} f_j(x) - \alpha_1 A_{j,i} a_i f_j(x) - \alpha_1 B_{j,i} f_j(x) \end{aligned}$$

have at most q roots over $\mathbb{F}_{q^t}$.

4. LINEAR SETS AND CYCLIC SUBSPACE CODES

In this section we first give a geometric description of the Sidon property in terms of linear sets. With the aid of Proposition 5 we then translate this description in terms of cyclic subspace codes.

Theorem 7. Let U be a k-dimensional $\mathbb{F}_q$-subspace of $\mathbb{F}_{q^n}$. Then U is a Sidon space if and only if the only points of $L_{U\times U} \subseteq \mathrm{PG}(1, q^n) = \mathrm{PG}(\mathbb{F}_{q^n} \times \mathbb{F}_{q^n}, \mathbb{F}_{q^n})$ of weight greater than one are those in $L_{U\times U} \cap \mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q)$. Furthermore, the weight of such points is k.

In particular, if U is a Sidon space then the size of $L_{U\times U}$ is

$$\frac{q^k-1}{q-1}(q^k - q) + q + 1.$$

Proof. Let $\alpha \in \mathbb{F}_{q^n}^*$. Let $\langle(1,\alpha)\rangle_{\mathbb{F}_{q^n}} \in L_{U\times U}$. Then there exists $\rho \in \mathbb{F}_{q^n}^*$ such that

$$\rho(1,\alpha) \in U \times U,$$

that is $\rho \in U \cap \alpha^{-1}U$. Therefore, if U is a Sidon space by Theorem 1 it follows that $\dim_{\mathbb{F}_q}(U \cap \alpha^{-1}U) \leq 1$ if $\alpha \notin \mathbb{F}_q$ and $\dim_{\mathbb{F}_q}(U \cap \alpha^{-1}U) = k$ if $\alpha \in \mathbb{F}_q$. So $w_{L_{U\times U}}(\langle(1,\alpha)\rangle_{\mathbb{F}_{q^n}}) = 1$ if and only if $\alpha \notin \mathbb{F}_q$ and if $w_{L_{U\times U}}(\langle(1,\alpha)\rangle_{\mathbb{F}_{q^n}}) \geq 2$ then $\alpha \in \mathbb{F}_q$ and $w_{L_{U\times U}}(\langle(1,\alpha)\rangle_{\mathbb{F}_{q^n}}) = k$. Suppose now that the only points of $L_{U\times U} \subseteq \mathrm{PG}(1, q^n)$ of weight greater than one are those in $L_{U\times U} \cap \mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q)$. Then if $\alpha \notin \mathbb{F}_q$ we have $\dim_{\mathbb{F}_q}(U \cap \alpha^{-1}U) \leq 1$ and so by Theorem 1 the subspace U turns out to be a Sidon space. The last part follows by 2 and 3.

The above result can be extended to the case of multi-Sidon spaces.

Theorem 8. Let $\{U_1, \dots, U_r\}$ be a set of $\mathbb{F}_q$-subspaces in $\mathbb{F}_{q^n}$ and let $k_i = \dim_{\mathbb{F}_q}(U_i)$ for every $i \in \{1, \dots, r\}$. Then $\{U_1, \dots, U_r\}$ is a multi-Sidon space if and only if

- the only points of $L_{U_i\times U_i} \subseteq \mathrm{PG}(1, q^n) = \mathrm{PG}(\mathbb{F}_{q^n} \times \mathbb{F}_{q^n}, \mathbb{F}_{q^n})$ of weight greater than one are those in $L_{U_i\times U_i} \cap \mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q)$, for every $i \in \{1, \dots, r\}$;
- $L_{U_i\times U_i} \cap L_{U_j\times U_j} = \mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q)$, for every $i, j \in \{1, \dots, r\}$ with $i \neq j$.

Proof. Let $\alpha \in \mathbb{F}_{q^n}^*$. Let $\langle(1,\alpha)\rangle_{\mathbb{F}_{q^n}} \in L_{U_i\times U_i} \cap L_{U_j\times U_j}$, with $i \neq j$. In particular, there exists $\rho \in \mathbb{F}_{q^n}^*$ such that

$$\rho(1,\alpha) = (u, \overline{u}),$$

for some $u, \overline{u} \in U_i$. Hence, $\alpha = \rho^{-1}\overline{u} = \overline{u}/u$. Similarly, $\alpha = \overline{v}/v$, for some $v, \overline{v} \in U_j$. So that $\alpha \in U_i \cdot U_i^{-1} \cap U_j \cdot U_j^{-1}$. The assertion now follows by (Zullo, 2021, Theorem 3.6) and by Theorem 7.

We can hence derive some bounds that involve the number and the dimensions of the subspaces of a multi-Sidon space, the degree of the field extension and q.

Theorem 9. Let $\{U_1, \dots, U_r\}$ be a multi-Sidon space and let $k_i = \dim_{\mathbb{F}_q}(U_i)$ for every $i \in \{1, \dots, r\}$. Then

$$\sum_{i=1}^{r} \frac{q^{k_i}-1}{q-1}(q^{k_i} - q) \leq q^n - q.$$

In particular, if $\{U_1, \dots, U_r\}$ is a maximum multi-Sidon space then

$$r \leq \frac{(q^n-q)(q-1)}{(q^{n/2}-q)(q^{n/2}-1)}.$$

If $n > 4$ then $r \leq q-1$ and if $n = 4$ then $r \leq q$.

Proof. By Theorem 8, the $L_{U_i \times U_i}$'s pairwise intersect each other only in $\mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q)$, so that

$$\left| \left(\bigcup_{i=1}^{r} L_{U_i \times U_i} \right) \setminus \mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q) \right| = \sum_{i=1}^{r} \frac{q^{k_i} - 1}{q - 1} (q^{k_i} - q).$$

Since $(\bigcup_{i=1}^{r} L_{U_i \times U_i}) \setminus \mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q) \subseteq \mathrm{PG}(1, q^n) \setminus \mathrm{PG}(\mathbb{F}_q \times \mathbb{F}_q, \mathbb{F}_q)$, the assertion follows.

Combining Proposition 5 and Theorem 9 we obtain the following bound on the number of orbits that a cyclic subspace codes (whose subspaces have dimension $n/2$) can have.

Corollary 10. Let $n = 2t$ and let $U_1, \ldots, U_r$ be $\mathbb{F}_q$-subspaces of dimension t in $\mathbb{F}_{q^n}$ and let

$$C = \bigcup_{i \in \{1,\ldots,r\}} C_{U_i} \subseteq \mathcal{G}_q(n, t)$$

be a subspace code. If the minimum distance of C is $2t-2$, then $r \leq q - 1$ if $n \geq 4$ and $r \leq q$ if $n = 4$.

We do not know if the above bound is tight but in (Roth et al., 2017, Construction 37) a construction of cyclic subspace codes as in Corollary 10 with r equals to roughly $q/2$ has been shown.

5. CONCLUSION

In this abstract we first give a generalization of the notion of Sidon space with the notion of multi-Sidon space. Then we show a link between multi-Sidon spaces of maximum dimension and cyclic subspace codes with certain parameters, which yields to a canonical form for such codes. Then we propose a geometric interpretation of the Sidon (and the multi-Sidon) property by means of linear sets, which give us an upper bound on the number of subspaces that a cyclic subspace codes associated with a multi-Sidon space can have. However, further combining the algebraic and geometric approaches, more results on cyclic subspace codes can be obtained. For instance, using the canonical form in Theorem 4, more examples of cyclic subspace codes could be shown.

ACKNOWLEDGEMENTS

I would like to thank Oriol Serra, who suggested me to study Sidon spaces during the conference *Discretaly* (2018).

REFERENCES

Bachoc, C., Serra, O., and Zémor, G. (2017). An analogue of vosper's theorem for extension fields. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 163, 423–452. Cambridge University Press.

Bachoc, C., Serra, O., and Zémor, G. (2018). Revisiting kneser's theorem for field extensions. *Combinatorica*, 38(4), 759–777.

Ben-Sasson, E., Etzion, T., Gabizon, A., and Raviv, N. (2016). Subspace polynomials and cyclic subspace codes. *IEEE Transactions on Information Theory*, 62(3), 1157–1165.

Etzion, T. and Vardy, A. (2011). Error-correcting codes in projective space. *IEEE Transactions on Information Theory*, 57(2), 1165–1173.

Hou, X.D., Leung, K.H., and Xiang, Q. (2002). A generalization of an addition theorem of kneser. *Journal of Number Theory*, 97(1), 1–9.

Koetter, R. and Kschischang, F.R. (2008). Coding for errors and erasures in random network coding. *IEEE Transactions on Information theory*, 54(8), 3579–3591.

Lavrauw, M. and Van de Voorde, G. (2015). Field reduction and linear sets in finite geometry. *Contemp. Math*, 632, 271–293.

Napolitano, V., Polverino, O., Santonastaso, P., and Zullo, F. (2021). Linear sets on the projective line with complementary weights. *arXiv preprint arXiv:2107.10641*.

O'Bryant, K. (2004). A complete annotated bibliography of work related to sidon sequences. *The Electronic Journal of Combinatorics*, 1000, DS11–Jul.

Otal, K. and Özbudak, F. (2017). Cyclic subspace codes via subspace polynomials. *Designs, Codes and Cryptography*, 85(2), 191–204.

Polverino, O. (2010). Linear sets in finite projective spaces. *Discrete mathematics*, 310(22), 3096–3107.

Roth, R.M., Raviv, N., and Tamo, I. (2017). Construction of sidon spaces with applications to coding. *IEEE Transactions on Information Theory*, 64(6), 4412–4422.

Trautmann, A.L., Manganiello, F., Braun, M., and Rosenthal, J. (2013). Cyclic orbit codes. *IEEE Transactions on Information Theory*, 59(11), 7386–7404.

Wu, B. and Liu, Z. (2013). Linearized polynomials over finite fields revisited. *Finite Fields and Their Applications*, 22, 79–100.

Zullo, F. (2021). Multi-sidon spaces over finite fields. *arXiv preprint arXiv:2112.08781*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Dissipation inequalities and sum-of-squares programming for reachability analysis

He Yin [*] **Peter Seiler** [**] **Murat Arcak** [***]

[*] *Department of Mechanical Engineering, University of California, Berkeley* he_yin@berkeley.edu.
[**] *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor* pseiler@umich.edu.
[***] *Department of Electrical Engineering and Computer Sciences, University of California, Berkeley* arcak@eecs.berkeley.edu.

Abstract: This abstract summarizes our recent results on reachability analysis using dissipation inequalities. We first outline a method to outer-approximate forward reachable sets (FRS) on finite horizons for uncertain polynomial systems. This method makes use of time-dependent polynomial storage functions that satisfy appropriate dissipation inequalities that account for $\mathcal{L}_2$ disturbances, uncertain parameters, and perturbations characterized by time-domain, integral quadratic constraints (IQC). By introducing IQCs to reachability analysis, we now allow for various types of uncertainty, including unmodeled dynamics. We next discuss backward reachable sets (BRS), and decompose control synthesis process into two steps: first we construct storage functions whose sublevel sets are used for BRS estimation, and then we compute control laws using these storage functions through quadratic programs (QP). In a separate result we simultaneously compute an under-approximation to the BRS, as well as an explicit control law in order to incorporate input saturation limits. These methods make use of the generalized S-procedure and Sum-of-Squares techniques to derive algorithms with the goal of finding the tightest approximation to the reachable sets.

Keywords: Nonlinear Systems and Control.

1. INTRODUCTION

Reachability analysis is of vital importance for safety-critical systems: it can verify whether a system is able to reach a target and avoid an obstacle. There are two fundamental types of reachable sets (Mitchell, 2007): forward and backward. The forward reachable set (FRS) is the set of all the successors of a given set of initial conditions under the system dynamics. The backward reachable set (BRS) is the set of initial conditions whose successors can be maintained safely inside a given state constraint set using an admissible control. In our works (Yin et al., 2018, 2019a,b), the forward and backward reachability problems are considered with finite horizons, since in many practical settings, systems only undergo finite-time trajectories. We make use of time-dependent storage functions that satisfy certain dissipation inequalities to characterize the FRS and BRS. The use of dissipation inequalities allows us to accommodate various sources of uncertainty, including $\mathcal{L}_2$ disturbances, uncertain parameters and perturbations Δ (e.g. unmodeled dynamics, uncertain time delay, and control saturation), whose input output properties are described by integral quadratic constraints (IQCs) (Megretski and Rantzer, 1997; Veenman et al., 2016). The generalized S-procedure and sum-of-squares (SOS) relaxation for polynomial non-negativity are used to derive computation algorithms for obtaining storage functions.

The work in (Yin et al., 2018) addresses the computation of an outer-approximation to the FRS by merging the dissipation inequality and IQCs. Therefore, although our nominal systems are assumed to be polynomials, including IQCs allows us to extend our analysis framework to non-polynomial systems. We formulate the FRS computation as generalized SOS optimization problems that are quasi-convex, which can be solved effectively by bisection, and for which global optimal solutions can be achieved. The work in (Yin et al., 2019a) studies backward reachability: we compute a storage function that characterizes an inner-approximation to the BRS first, and then we obtain the min-norm control law as the closed-form solution to the quadratic programming (QP) based on the computed storage function. The control law is not restricted to polynomial functions. In (Yin et al., 2019b), the BRS inner-approximation and control law are computed at the same time in order to account for control saturation. In these two works, the derived optimizations are nonconvex due to bilinearity in decision variables. Therefore, algorithms are designed to alternate the search over bilinear variables.

In the existing literature, there are various approaches to reachability analysis, including polytopic methods (Borrelli et al., 2011), Hamilton-Jacobi methods (Mitchell and Tomlin, 2000), ellipsoid methods (Kurzhanskiy and Varaiya, 2007) and interval analysis (Jaulin et al., 2001). Finite horizon forward reachability analysis is also considered in (Majumdar and Tedrake, 2017), but it only allows parametric uncertainty. The BRS is *outer-approximated* in (Henrion and Korda, 2014) by taking the complement of

the initial set from which no trajectory is able to reach the target set for any admissible inputs. The result in (Henrion and Korda, 2014) is complementary to our works (Yin et al., 2019a,b), since we provide inner-approximations to the BRS, as well as an explict control law.

The abstract is organized as follows. In Section 2, the method for forward reachability analysis from (Yin et al., 2018) is described. In Section 3.1, the QP based control synthesis method from (Yin et al., 2019a) is discussed. In Section 3.2, the method from (Yin et al., 2019b) that synthesizes the storage function and control law at the same time is summarized. Section 4 provides the way of formulating SOS problems. Section 5 summarizes the results and gives possible directions for the future work.

2. FORWARD REACHABILITY ANALYSIS

The proposed forward reachability analysis framework considers the following uncertain nonlinear system:

$$\dot{x} = f(t, x, w, l), \tag{1a}$$
$$v = h(t, x, w, l), \tag{1b}$$
$$l = \Delta(v), \tag{1c}$$

with $x \in X \subset \mathbb{R}^n$, $l \in \mathbb{R}^{n_l}$, $w \in \mathbb{R}^{n_w}$, $v \in \mathbb{R}^{n_v}$, where f, h define the nominal system G and the perturbation Δ is an operator $\Delta : \mathcal{L}_2^{n_v}[0,T] \to \mathcal{L}_2^{n_l}[0,T]$. The uncertain system (1), denoted as $F_u(G,\Delta)$, is an interconnection of Δ and G, as shown in Fig. 1.

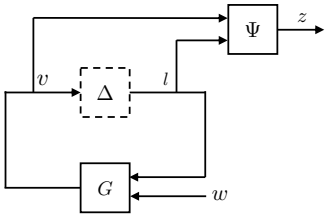


Fig. 1. Interconnection $F_u(G,\Delta)$ of a nominal system G and a perturbation Δ

Assumption 1. (i) the disturbance w has bounded $\mathcal{L}_2$ energy: $\int_0^T w^\top(t)w(t)dt \le R^2$, with R given, and (ii) all the trajectories $x(t)$ of $F_u(G,\Delta)$ start in the set $X_0 \subset X$.

Under this assumption, let $FRS(T; X_0) := \{x(T) \mid x(t) \in X_0\}$ denote the FRS of $F_u(G,\Delta)$. Our goal is to outer-approximate this $FRS(T; X_0)$.

The perturbation Δ is characterized by IQCs. To define an IQC we introduce a virtual Ψ (shown in Fig. 1) that is an LTI system, driven by the input and output of Δ, and with zero initial condition $x_\psi(0) = 0$. The dynamics of Ψ are given by

$$\dot{x}_\psi = A_\psi x_\psi + B_{\psi 1} v + B_{\psi 2} l, \tag{2a}$$
$$z = C_\psi x_\psi + D_{\psi 1} v + D_{\psi 2} l, \tag{2b}$$

where $x_\psi \in \mathbb{R}^{n_\psi}$, and $z \in \mathbb{R}^{n_z}$. A hard IQC is a quadratic constraint enforced on the output z of Ψ associated with a matrix $M \in \mathbb{S}^{n_z}$ over all finite horizons. Δ is said to satisfy the hard IQC defined by (Ψ, M), if for all $v \in \mathcal{L}_2^{n_v}[0,T]$, $l = \Delta(v)$, we have

$$\int_0^t z(\tau)^\top M z(\tau) d\tau \ge 0, \ \forall t \in [0,T]. \tag{3}$$

The notation $\Delta \in \text{HardIQC}(\Psi, M)$ is used to indicate that Δ satisfies the corresponding hard IQC. Therefore, a perturbation can be replaced by a filter Ψ and an IQC (3). The outer-approximation to the FRS of $F_u(G,\Delta)$ is then computed using an extended system, which is an interconnection of G and Ψ with its corresponding IQC (3). Assume M is constrained to a convex set $\mathcal{M}$, which is described by linear matrix inequalities (LMIs). For more details on $\mathcal{M}$, the reader is referred to (Veenman et al., 2016). The following constraints, including dissipation inequality (4a), provide the analysis conditions for $F_u(G,\Delta)$ with $\Delta \in$ IQC(Ψ, M):

$$\partial_t V + \partial_x V \cdot f(t,x,w,l) + \partial_{x_\psi} V \cdot (A_\psi x_\psi + B_{\psi 1} v + B_{\psi 2} l) + z^\top M z \le w^\top w, \quad \forall (t, x, x_\psi, l, w) \in [0,T] \times X \times \mathbb{R}^{n_\psi} \times \mathbb{R}^{n_l} \times \mathbb{R}^{n_w}, \tag{4a}$$
$$X_0 \subseteq \{x : V(0,x,0) \le 0\}, \tag{4b}$$
$$\{x : V(t,x,x_\psi) \le R^2\} \subseteq X, \forall (t, x_\psi) \in [0,T] \times \mathbb{R}^{n_\psi}, \tag{4c}$$
$$M \in \mathcal{M}, \tag{4d}$$

where $V(t,x,x_\psi)$ and M are decision variables, while (G,Ψ) and (T,R,X,X_0) are fixed. The set X is the region, over which the dissipation inequality holds. If constraints (4) hold, then $\{(x,x_\psi) : V(T,x,x_\psi) \le R^2\}$ outer-approximates the FRS of the extended system (G,Ψ) at the terminal time T from the initial set $X_0 \times 0$. Finally the projection of $\{(x,x_\psi) : V(T,x,x_\psi) \le R^2\}$ on the x space outer-approximates the FRS of $F_u(G,\Delta)$.

3. BACKWARD REACHABILITY ANALYSIS AND SYNTHESIS

The backward reachability analysis and synthesis framework focuses on nonlinear systems of the form

$$\dot{x} = f(t,x) + g(t,x)u + g_w(t,x)w, \tag{5}$$

where the control input $u \in \mathbb{R}^{n_u}$ and the external disturbance $w \in \mathbb{R}^{n_w}$ enter the system affinely. Assume that we are given a target set $X_T \subset \mathbb{R}^n$ for the system (5) to reach at the terminal time T. Assume (i) w has bounded $\mathcal{L}_2$ energy: $\int_0^T w^\top(t)w(t)dt \le R^2$, with R given, (ii) w satisfies an $\mathcal{L}_\infty$ constraint: $w(t) \in W := \{\eta \in \mathbb{R}^{n_w} : \eta^\top \eta \le \alpha\}$, for all $t \in [0,T]$. Let $x(t; x_0, u)$ define the solution to the system (5) at time t, from the initial condition $x(0) = x_0$ under the control u. The BRS is defined as $BRS(T, X_T) := \{\xi \in \mathbb{R}^n : \exists u(\cdot), \text{ s.t. } x(T;\xi,u) \in X_T\}$. The goal is to inner-approximate the BRS, and to find a control law that is able to steer all the trajectories initialized from the inner-approximation to X_T.

3.1 QP based Synthesis

In this section, we describe the method presented in (Yin et al., 2019a), where the computation of BRS estimation and control law is decomposed into two steps. The first step is to compute the storage function that characterizes the BRS inner-approximation, and the second is to solve for the minimum-norm control law using a QP that involves the obtained storage function. The following constraints characterize the storage function to be found in the first step:

$$\partial_t V + \partial_x V \cdot (f(t,x) + g_w(t,x)w) \leq w^\top w, \forall (t,x,w) \text{ s.t.}$$
$$\partial_x V \cdot g(t,x) = 0,\ V(t,x) \leq R^2,\ t \in [0,T],\ w \in W, \quad (6a)$$
$$\{x : V(T,x) \leq R^2\} \subseteq X_T, \quad (6b)$$

where (f,g,g_w,R,T,W,X_T) are fixed, while $V(t,x)$ is the decision variable. (6a) implies that we can always find a u of proper sign and sufficiently large magnitude such that the following dissipation inequality holds

$$\partial_t V + \partial_x V \cdot (f(t,x) + g(t,x)u(t) + g_w(t,x)w) \leq w^\top w,$$
$$\forall (t,x,w) \text{ s.t. } V(t,x) \leq R^2, t \in [0,T], w \in W.$$

Therefore, if (6) holds, then there exists a control $u(\cdot)$, such that $x(T) \in X_T$ for all $x(0) \in \{x : V(0,x) \leq 0\}$, in other words, the set $\{x : V(0,x) \leq 0\}$ is an inner-approximation to the BRS.

After V is obtained, we look for the input $u(t)$ that ensures the satisfaction of the dissipation inequality, with the magnitude of the input been minimized. To achieve it, the min-nrom input $u^\star(t)$ is given as the closed-form solution to the following QP:

$$\min_{u \in \mathbb{R}^{n_u}} u^\top u \quad (7a)$$
$$\text{s.t. } \max_{w \in W} \{\partial_t V + \partial_x V \cdot (f + gu + g_w w) \leq w^\top w\}. \quad (7b)$$

Since (7b) is quadratic in w, we can solve for the worst-case disturbance $w^\star$ using KKT conditions, and obtain $w^\star(t,x) =$

$$\begin{cases} \dfrac{\sqrt{\alpha}}{\sqrt{c(t,x)^\top c(t,x)}} c(t,x), & \text{if } c(t,x)^\top c(t,x) \geq 4\alpha, \\ \dfrac{1}{2} c(t,x), & \text{else,} \end{cases} \quad (8)$$

where $c := (\partial_x V \cdot g_w)^\top$. Substituting (8) back into (7b) yields two QPs corresponding to two cases listed in (8). Solving both QPs using KKT conditions gives the explicit expressions for the min-norm control input

$$u^\star(t,x) = \begin{cases} 0, & \text{if } b(t,x) \leq 0, \\ \dfrac{-b(t,x)}{a(t,x)a(t,x)^\top} a(t,x)^\top & \text{else,} \end{cases}$$

where $a := \partial_x V \cdot g$ and

$$b := \begin{cases} \partial_t V + \partial_x V \cdot f + \sqrt{\alpha c^\top c} - \alpha, & \text{if } c^\top c \geq 4\alpha \\ \partial_t V + \partial_x V \cdot f + c^\top c/4 - \alpha, & \text{else.} \end{cases}$$

An advantage of this method is that given a storage function, explicit control laws with smallest feasible magnitude can be derived, and they are not restricted to polynomials. However, control saturation is not considered in this framework.

3.2 $V - k$ iterative Synthesis

In this section, we summarize the method in (Yin et al., 2019b), where we search for a storage function V and a control law k at the same time so as to take control saturation into account. Let $k : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n_u}$ define a memoryless time-varying state feedback control by $u(t) = k(t, x(t))$. Assume the set of control constraints is given as a polytope $U = \{u \in \mathbb{R}^{n_u} : Au \leq h\}$, where $A \in \mathbb{R}^{n_c \times m}$ and $h \in \mathbb{R}^{n_c}$. Now the goal for the controller is to steer system (5) from the BRS inner-approximation to X_T at time T, while satisfying $u(t) \in U$ for all $t \in [0,T]$. Here we provide the sufficient conditions for V to characterize inner-approximations and k to achieve the control objective:

$$\partial_t V + \partial_x V \cdot (f(t,x) + g(t,x)k(t,x) + g_w(t,x)w) \leq w^\top w,$$
$$\forall (t,x,w) \text{ s.t. } V(t,x) \leq R^2, t \in [0,T], w \in W, \quad (9a)$$
$$\{x : V(T,x) \leq R^2\} \subseteq X_T, \quad (9b)$$
$$Ak(t,x) \leq h,\ \forall (t,x) \text{ s.t. } V(t,x) \leq R^2, t \in [0,T], \quad (9c)$$

where $V(t,x)$ and $k(t,x)$ are decision variables. If constraints (9) hold, then for all $x(0) \in \{x : V(0,x) \leq 0\}$, we guarantee $x(T) \in X_T$, under control law k. The main idea behind the constraint (9c) is that while a state $x(t)$ stays in the set $\{x : V(t,x) \leq R^2\}$, the control input derived from $u(t) = k(t, x(t))$ satisfies the control saturation.

4. SUM-OF-SQUARES FORMULATIONS

In general, looking for a generic function V that satisfies (4) or (6), and (V,k) that satisfy (9) could be difficult. Therefore, we use sum-of-squares programming in finding those decision variables. Notice that (4), (6) and (9) are all set-containment constraints, which can be certified by the generalized S-procedure, along with a method to check non-negativity. SOS relaxations can be used in checking non-negativity if all functions are restricted to polynomials. Therefore, we restrict the system model, control law and storage function to be polynomials, and we assume all the sets X_0, X_T, X are sublevel sets of polynomials. Now we are ready to derive SOS optimization problems. Take (9a) for example, if the following conditions hold, then (9a) is satisfied: s_a, s_b, s_c are SOS polynomials, and

$$-(\dot{V} - w^\top w) + (V - R^2)s_a - t(T-t)s_b - (\alpha - w^\top w)s_c$$

is an SOS polynomial, where decision variables s_a, s_b, and s_c are called S-procedure certificate. Checking if a polynomial is an SOS polynomial can be done by solving a corresponding semidefinite programming (Parrilo, 2000). Convex SOS problem can be derived for (4). However, (6) and (9) result in nonconvex SOS problems, due to the bilinearity in V, and (k, s_i). But we can still tackle these nonconvex problems by alternating the search over V and (k, s_i).

5. CONCLUSIONS AND FUTURE WORK

This extended abstract summarizes the works from (Yin et al., 2018, 2019a,b), where dissipation inequality based methods are proposed to approximate the reachable sets and synthesize control laws. In (Yin et al., 2018), the computation of an outer-approximation to the FRS is proposed, where IQCs are incorporated to model a large variety of uncertainties. In (Yin et al., 2019a), a storage function that characterizes an inner-approximation to the BRS is computed first, and a min-norm control law is obtained by solving a QP that involves the computed storage function. Finally, in (Yin et al., 2019b), a BRS inner-approximation and control law are computed simultaneously to consider input saturation limits. In both (Yin et al., 2019a,b), $\mathcal{L}_2$ disturbances and uncertain parameters are considered.

As for future work, we will continue to improve the computational efficiency and scalability of the reachability algorithms that are currently being developed. In addition we will combine these algorithms with data-driven methods

for estimating reachable sets with probabilistic guarantees. We will develop a formal definition of a data-driven reachable set approximation that is correct in a probabilistic sense, and devise appropriate sampling schemes to generate such approximations.

ACKNOWLEDGEMENTS

This work is funded in part by the grants ONR grant N00014-18-1-2209, AFOSR FA9550-18-1-0253, and NSF ECCS-1906164.

REFERENCES

Borrelli, F., Bemporad, A., and Morari, M. (2011). *Predictive control for linear and hybrid systems*. Cambridge University Press.

Henrion, D. and Korda, M. (2014). Convex computation of the region of attraction of polynomial control systems. *IEEE Transactions on Automatic Control*, 59, 297–312.

Jaulin, L., Kieffer, M., Didrit, O., and Walter, E. (2001). *Applied interval analysis: with examples in parameter and state estimation, robust control and robotics*, volume 1. Springer Science & Business Media.

Kurzhanskiy, A. and Varaiya, P. (2007). Ellipsoidal techniques for reachability analysis of discrete-time linear systems. *IEEE TAC*, 52, 26–38.

Majumdar, A. and Tedrake, R. (2017). Funnel libraries for real-time robust feedback motion planning. *The international Journal of Robotics Research*, 36, 947–982.

Megretski, A. and Rantzer, A. (1997). System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42, 819–830.

Mitchell, I. and Tomlin, C. (2000). Level set methods for computation in hybrid systems. In *In Hybrid Systems: Computation and Control*, 310–323.

Mitchell, I.M. (2007). Comparing forward and backward reachability as tools for safety analysis. In A. Bemporad, A. Bicchi, and G. Buttazzo (eds.), *Hybrid Systems: Computation and Control*, 428–443. Springer Berlin Heidelberg, Berlin, Heidelberg.

Parrilo, P. (2000). Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. PhD thesis, California Institute of Technology.

Veenman, J., Scherer, C., and Koroglu, H. (2016). Robust stability and performance analysis based on integral quadratic constraints. *European Journal of Control*, 31, 1–32.

Yin, H., Packard, A., Arcak, M., and Seiler, P. (2019a). Finite horizon backward reachability analysis and control synthesis for uncertain nonlinear systems. In *2019 American Control Conference (ACC)*, 5020–5026. doi: 10.23919/ACC.2019.8814444.

Yin, H., Arcak, M., Packard, A., and Seiler, P. (2019b). Backward Reachability for Polynomial Systems on A Finite Horizon. *arXiv e-prints*, arXiv:1907.03225.

Yin, H., Packard, A., Arcak, M., and Seiler, P. (2018). Reachability Analysis Using Dissipation Inequalities For Nonlinear Dynamical Systems. *arXiv e-prints*, arXiv:1808.02585.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Flat outputs for funnel control of non-minimum-phase systems

Thomas Berger * **Timo Reis** ** **Leonie Wagner** ***

* *Fachbereich Mathematik, Universität Paderborn, Technologiepark 21, 33100 Paderborn, Germany (e-mail: thomas.berger@upb.de).*
** *Institut für Mathematik, Technische Universität Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (e-mail: timo.reis@tu-ilmenau.de).*
*** *Fachbereich Mathematik, Universität Hamburg, Bundesstraße 55, 22083 Hamburg, Germany (e-mail: leonie.wagner@gmx.de).*

Abstract: We consider adaptive ouput feedback tracking control of linear time-invariant systems which are not necessarily minimum phase. The zero dynamics is split into a stable and an unstable part, we show that a flat output of the unstable part can contribute to the design of a funnel controller of the system. More precisely, we consider an auxiliary output based of the "true output" of the system and the flat output of the unstable part of the zero dynamics. The funnel controller is designed for this auxiliary output, and the consequences for the true output are discussed.

Keywords: linear systems; robust control; non-minimum phase; funnel control; relative degree

1. SYSTEM CLASS

We consider stabilizable linear systems given by

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad x(0) = x^0 \in \mathbb{R}^n, \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B, C^\top \in \mathbb{R}^{n\times m}$, with the same number of inputs $u : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ and outputs $y : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$. We assume that (1) has strict relative degree $r \in \mathbb{N}$, that is

$$CA^kB = 0,\ k = 0, \ldots, r-2,\ CA^{r-1}B \in \mathrm{Gl}_n(\mathbb{R}), \tag{2}$$

cf. Isidori (1995). While adaptive control of minimum phase linear systems is well-studied, see e.g. the classical works Byrnes and Willems (1984); Khalil and Saberi (1987); Morse (1983), we stress that we do not assume that (1) is minimum phase or, equivalently, its zero dynamics are asymptotically stable. The latter would mean that $\mathrm{rk}\left[\begin{smallmatrix} A-\lambda I_n & B \\ C & 0 \end{smallmatrix}\right] = n + m$ for all $\lambda \in \mathbb{C}_-$, see e.g. Ilchmann et al. (2007); Isidori (1995). By the Byrnes-Isidori form (Ilchmann et al., 2007, Lem. 3.5) (see also Isidori (1995)) we have that, if (2) is satisfied, there exists a state-space transformation $U \in \mathrm{Gl}_n(\mathbb{R})$ such that $Ux(t) = \big(y(t)^\top, \dot{y}(t)^\top, \ldots, y^{(r-1)}(t)^\top, \eta(t)^\top\big)^\top$, where $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n-rm}$, transforms (1) into

$$\begin{aligned} y^{(r)}(t) &= \sum_{i=1}^{r} R_i y^{(i-1)}(t) + S\eta(t) + \Gamma u(t), \\ \dot{\eta}(t) &= Py(t) + Q\eta(t) + d_\eta(t), \end{aligned} \tag{3}$$

where $R_i \in \mathbb{R}^{m\times m}$ for $i = 1, \ldots, r$, $S, P^\top \in \mathbb{R}^{m\times(n-rm)}$, $Q \in \mathbb{R}^{(n-rm)\times(n-rm)}$, and $\Gamma := CA^{r-1}B$. Furthermore, (1) is minimum phase if, and only if, $\sigma(Q) \subseteq \mathbb{C}_-$. The second equation in (3) represents the internal dynamics of the linear system (1); if $y = 0$, then these dynamics are called zero dynamics.

Our assumption is that the system does not have any zeros on the imaginary axis. Consequently, there exists $T \in \mathrm{Gl}_{n-rm}(\mathbb{R})$ and $\ell \in \mathbb{N}$ such that

$$TQT^{-1} = \begin{bmatrix} \hat{Q}_1 & \hat{Q}_2 \\ 0 & \tilde{Q} \end{bmatrix}, \quad TP = \begin{bmatrix} \hat{P} \\ \tilde{P} \end{bmatrix}, \tag{4}$$

where $\hat{Q}_1 \in \mathbb{R}^{(n-rm-\ell)\times(n-rm-\ell)}$, $\hat{Q}_2 \in \mathbb{R}^{(n-rm-\ell)\times\ell}$, $\tilde{Q} \in \mathbb{R}^{\ell\times\ell}$, $\hat{P} \in \mathbb{R}^{(n-rm-\ell)\times m}$. $\tilde{P} \in \mathbb{R}^{\ell\times m}$ with $\sigma(\hat{Q}_1) \subseteq \mathbb{C}_-$, and $(\tilde{Q}, \tilde{P})$ is controllable.

2. CONTROL OBJECTIVE

To treat the non-minimum phase property of system (1) the system parameters A, B, C need to be known, at least partially, and additional components of the state x need to be available to the controller. For the time being, assume that the measurement of a partial state $\hat{x}(t) = Hx(t)$ is available, where H will be specified by the presented controller design. We stress that the measurement of the full state $x(\cdot)$ or knowledge of the full initial value x^0 is, in general, not required. Therefore, the objective is to design a dynamic partial state feedback of the form

$$\begin{aligned} \dot{z}(t) &= F\big(t, z(t), \hat{x}(t), y_{\mathrm{ref}}(t)\big), \quad z(0) = z^0, \\ u(t) &= G\big(t, z(t), \hat{x}(t), y_{\mathrm{ref}}(t)\big), \end{aligned} \tag{5}$$

where $y_{\mathrm{ref}} : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ is a sufficiently smooth reference signal, such that in the closed-loop system the tracking error $e(t) = y(t) - y_{\mathrm{ref}}(t)$ evolves within a prescribed performance funnel

$$\mathcal{F}_\varphi := \{\ (t, e) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^m \ \mid \varphi(t)\|e\| < 1\ \}, \tag{6}$$

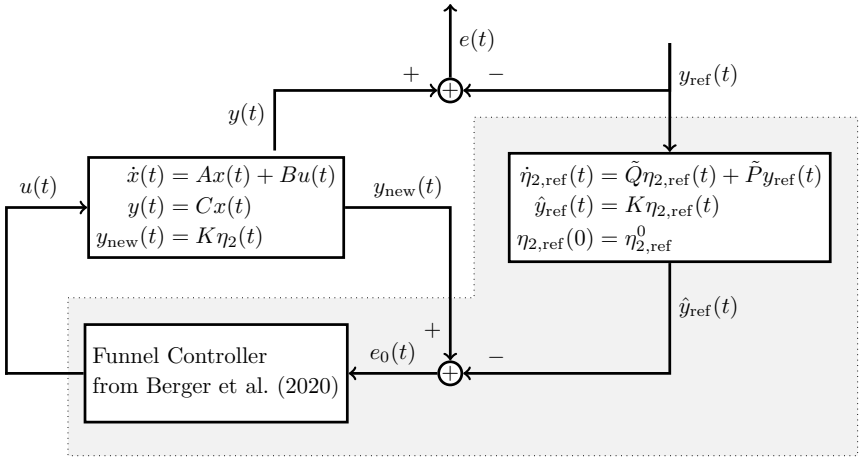


Fig. 1. The funnel controller, indicated by the grey box, applied to system (1) with new output as in (8). The controller consists of the generator of the new reference signal (9) and the funnel controller developed in Berger et al. (2020).

which is determined by a positive function $\varphi \in \mathcal{C}^r(\mathbb{R}_{\geq 0} \to \mathbb{R})$ with bounded $\varphi, \dot{\varphi}, \ldots, \varphi^{(r)}$, and $\liminf_{\tau\to\infty} \varphi(\tau) > 0$.

Furthermore, all signals x, u, z should remain bounded, even though (1) is non-minimum phase.

The funnel boundary is given by the reciprocal of φ as depicted in Fig. 2. Each performance funnel with φ as

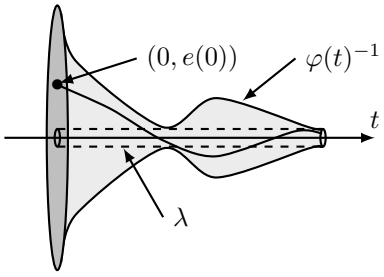


Fig. 2. Error evolution in a funnel with boundary $\varphi(t)^{-1}$.

above is bounded away from zero.

3. THE CONTROLLER

With the decomposition of Q as in (4) we may further transform the system from (3) using $T\eta = (\eta_1^\top, \eta_2^\top)^\top$ with $\eta_1 : \mathbb{R}_{\geq 0} \to \mathbb{R}^{n-rm-\ell}$, $\eta_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}^\ell$ into

$$\begin{aligned} y^{(r)}(t) &= \sum_{i=1}^{r} R_i y^{(i-1)}(t) + S_1\eta_1(t) + S_2\eta_2(t) + \Gamma u(t), \\ \dot{\eta}_1(t) &= \hat{Q}_1\eta_1(t) + \hat{Q}_2\eta_2(t) + \hat{P}y(t), \\ \dot{\eta}_2(t) &= \tilde{Q}\eta_2(t) + \tilde{P}y(t), \end{aligned} \tag{7}$$

where $[S_1, S_2] = ST^{-1}$. Since $(\tilde{Q}, \tilde{P})$ is controllable, it is possible to construct an artificial output

$$y_{\text{new}}(t) := K\eta_2(t), \tag{8}$$

which is flat in the sense of Fliess et al. (1995). It can be shown that such an output can be chosen in a way that the system with input u and output y_{new} has a well-defined vector relative degree in the sense of Berger et al. (2020); Mueller (2009). This enables to apply the funnel controller from Berger et al. (2020) for systems with vector relative degree. By the construction of the new output in (8), the new reference signal is generated by the corresponding subsystem of (7) when the original reference signal is inserted for the original output, i.e.,

$$\begin{aligned} \dot{\eta}_{2,\text{ref}}(t) &= \tilde{Q}\eta_{2,\text{ref}}(t) + \tilde{P}y_{\text{ref}}(t), \quad \eta_{2,\text{ref}}(0) = \eta^0_{2,\text{ref}}, \\ \hat{y}_{\text{ref}}(t) &= K\eta_{2,\text{ref}}(t). \end{aligned} \tag{9}$$

The overall controller structure is depicted in Fig. 1.

We show that this controller results in a global and bounded solution, and we will discuss the performance of the output $y(\cdot)$ and its distance to the reference trajectory y_{ref}.

REFERENCES

Berger, T. (2020). Tracking with prescribed performance for linear non-minimum phase systems. *Automatica*, 115, Article 108909.

Berger, T., Lê, H.H., and Reis, T. (2020). Vector relative degree and funnel control for differential-algebraic systems. In S. Grundel, T. Reis, and S. Schöps (eds.), *Progress in Differential-Algebraic Equations II*, Differential-Algebraic Equations Forum, 213–255. Springer, Berlin-Heidelberg.

Byrnes, C.I. and Willems, J.C. (1984). Adaptive stabilization of multivariable linear systems. In *Proc. 23rd IEEE Conf. Decis. Control*, 1574–1577.

Fliess, M., Levine, J., Martin, P., and Rouchon, P. (1995). Flatness and defect of non-linear-systems: introductory theory and examples. *Int. J. Control*, 61, 1327–1361.

Ilchmann, A., Ryan, E.P., and Townsend, P. (2007). Tracking with prescribed transient behavior for nonlinear systems of known relative degree. *SIAM J. Control Optim.*, 46(1), 210–230.

Isidori, A. (1995). *Nonlinear Control Systems*. Communications and Control Engineering Series. Springer-Verlag, Berlin, 3rd edition.

Khalil, H.K. and Saberi, A. (1987). Adaptive stabilization of a class of nonlinear systems using high-gain feedback. *IEEE Trans. Autom. Control*, 32, 1031–1035.

Morse, A.S. (1983). Recent problems in parameter adaptive control. In I.D. Landau (ed.), *Outils et Modèles Mathématiques pour l'Automatique, l'Analyse de Systèmes et le Traitment du Signal*, volume 3, 733–740. Éditions du Centre National de la Recherche Scientifique (CNRS), Paris.

Mueller, M. (2009). Normal form for linear systems with respect to its vector relative degree. *Linear Algebra Appl.*, 430(4), 1292–1312.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Stochastic optimal control for nonlinear damped serial network dynamics

Simone Göttlich * **Thomas Schillinger** *

* *University of Mannheim, Department of Mathematics, 68131 Mannheim, Germany (e-mail: goettlich@uni-mannheim.de, schillinger@uni-mannheim.de)*

Abstract: We present a stochastic optimal control problem for a serial network. The dynamics of the network are governed by transport equations with a special emphasis on the nonlinear damping function. The demand profile at the network sink is modeled by a stochastic differential equation. An explicit optimal inflow into the network is determined and numerical simulations are presented to show the effects for different choices of the nonlinear damping.

Keywords: Optimal stochastic control, uncertain demands, transport equations, nonlinear damping
AMS Classification: 93E20, 65C20, 60H10

1. INTRODUCTION

Energy networks or supply networks in a more general setting can be analyzed from different perspectives. Mathematically, there is a strong interest to describe the dynamics inside the network to better understand the underlying physical or economic processes. Typical applications range from electric transmission lines Göttlich et al. (2015) and gas networks Banda et al. (2006) to production systems D'Apice et al. (2010), wherein the dynamics is governed by nonlinear hyperbolic transport equations Bressan et al. (2014). Generally, nonlinearities might appear due to complex flow patterns or additional interaction terms such as resistance or friction. For many applications the question arises how such a network can be controlled to satisfy consumer demands. Since demands usually include a kind of uncertainty, this leads to a challenging stochastic optimal control problem. To reduce the complexity of the problem, we therefore start with the consideration of linearized dynamics and nonlinear damping combined with stochasticity of demands in an abstract setting.

This work is based on ideas originally presented in Göttlich and Schillinger (2021); Göttlich et al. (2019). Basically, we consider a network framework consisting of three ingredients. At the network source an optimal inflow shall be injected into the system such that a given demand is met. On every arc of the network transport equations of hyperbolic type describe the dynamics. In our case we focus on linear flux functions and add unlike Göttlich and Schillinger (2021) a nonlinear damping term. However, we are still able to compute an explicit representation of the optimal input. As a last component, we consider uncertain demands at the sinks of the network. These are described by a stochastic process, given by a stochastic differential equation. In this work, we assume that demands are described by Jacobi processes, as recently proposed by Coskun and Korn (2021). These are mean-reverting processes which stay in a bounded interval and are therefore very suitable for various applications we have in mind. The goal of this contribution is to show under which assumptions we will be able to explicitly derive the optimal inflow subject to the network dynamics and the stochastic demand.

The organization of the article is the following. In Section 2 we present the full optimal control framework. The discussion of the objective function and the availability of information is executed in Section 3. The key result is presented in Section 4, where we extensively describe how an explicit formula for the optimal inflow can be calculated in the setting of nonlinear damping terms. Section 5 concludes with a numerical study of different nonlinear damping functions.

2. THE OPTIMAL CONTROL PROBLEM

We consider the control problem in (1), where we aim to determine the optimal input $u(t)$ given some stochastic demand D_t and the network dynamics in terms of scalar transport equations. More precisely, we make use of the following notation: For a directed serial network, the set C denotes the demand node and J the set of the inner nodes. We control the network inflow $u(t)$ defined in (1d) and solve an optimization problem for the demand node. At the demand node we consider a demand process $(D_t)_{t\in[t_0,T]}$ for which we assume a Jacobi process. A main advantage of the Jacobi process apart from the mean-reverting behaviour is its boundedness such that negative or arbitrarily large demands cannot be attained. It can be described by a stochastic differential equation given in (1f)-(1g), where θ is a time-dependent mean reversion level, κ the mean reversion speed and σ the scaling of the stochastic disturbances coming from a Brownian motion $(W_t)_{t\in[t_0,T]}$. In this work, we only consider Jacobi processes on $[0,1]$, which can be easily translated on any bounded interval.

$$\min_{\substack{u\in L^2\\ i:v_i\in C}} \int_{t_0}^{T} \mathbb{E}\left[\left(D_s - f^{(i)}(z^{(i)}(1,s),s)\right)^2 \;\middle|\; \mathcal{F}_{\hat{t}}\right] ds \quad \text{(1a)}$$

$$\text{s.t. } z_t^{(i)}(x,t) + f^{(i)}(z^{(i)}(x,t),t)_x + g^{(i)}(z^{(i)}(x,t),t) = 0 \tag{1b}$$

$$z^{(i)}(x,t_0) = z_0^{(i)}(x), \quad \forall i \text{ s.t. } v_i \in J \cup C \tag{1c}$$

$$f^{(1)}(z^{(1)}(0,t),t) = u(t) \tag{1d}$$

$$f^{(i+1)}(z^{(i+1)}(0,t),t) = f^{(i)}(z^{(i)}(1,t),t) \quad \forall i \in J \tag{1e}$$

$$dD_t = \kappa\left(\theta(t) - D_t\right) + \sigma\sqrt{D_t\left(1-D_t\right)}dW_t \tag{1f}$$

$$D_0 = d_0. \tag{1g}$$

We aim to minimize the expected quadratic deviation between the demand process and the outflow of the network with respect to a filtration $(\mathcal{F}_t, t \geq 0)$, see (1a). In the objective function, we condition on a time $\hat{t} \leq t_0$ up to which demand information is available. Section 3 presents further strategies about demand updates. On the network arcs i we consider hyperbolic partial differential equations for the quantities $z^{(i)}$ in (1b), where the flux functions $f^{(i)}$ are linear with respect to $z^{(i)}$ and possibly nonlinear in time, accompanied with some initial data (1c). In comparison to the work in Göttlich and Schillinger (2021), the damping functions $g^{(i)}$ are allowed the be nonlinear in the quantities $z^{(i)}$. The choices of the damping functions will be discussed in Section 4. At each inner node due to the serial network structure, there is exactly one ingoing and one outgoing arc at which we ensure flux conservation (1e).

3. SOLUTION TO THE OBJECTIVE FUNCTION AND DEMAND UPDATES

The optimal control problem in (1) has a stochastic component given by the stochastic differential equation (SDE) for the demand process (1f). But when minimizing the expected quadratic deviation of demand and network supply the optimal solution for a square integrable demand process is given by $\mathbb{E}[D_t|\mathcal{F}_{\hat{t}}]$ (see Corollary 8.17 in Klenke (2020)). Therefore, the particular structure of the demand process, apart from the first two conditional moments, does not matter to solve the optimal control problem. Their explicit forms in case of the Jacobi process can e.g. be found in Delbaen and Shirakawa (2002). As a next step, we want to generalize the control problem a little further. So far, we have assumed that there is one time $\hat{t} \leq t_0$ at which the demand levels are updated. However, we now consider a sequence of update times $(\hat{t}_j)_{j\in\mathbb{N}}$ which allow for additional information about the demands. Then, we end up with a sequence of optimal control problems on time intervals $[\hat{t}_j, \hat{t}_{j+1})$. To account for the correct time intervals, the objective function from (1a) then reads

$$\min_{\substack{u\in L^2 \\ i:v_i\in C}} \int_{\tilde{t}(v_i,\hat{t}_j)}^{\tilde{t}(v_i,\hat{t}_{j+1})} \mathbb{E}\Bigg[\Bigg(D_s - f^{(i)}(z^{(i)}(1,s),s)\Bigg)^2 \;\Big|\; \mathcal{F}_{\hat{t}_j}\Bigg] ds,$$

where $\tilde{t}(v_i,t)$ denotes the time at which the inflow inserted at the source node at time t reaches the demand node v_i. Additionally, for all updates except the first one, the initial data has to correspond to the state of the system of the time interval before. Therefore, the equations (1c) and (1g) read

$$z^{(i)}(x,\hat{t}_j) = z_{\text{old}}^{(i)}(x,\hat{t}_j), \quad \forall i \text{ s.t. } v_i \in J \cup C$$
$$D_{\hat{t}_j} = D_{\hat{t}_j}^{\text{old}},$$

where $z_{\text{old}}^{(i)}(x,\hat{t}_j)$ and $D_{\hat{t}_j}^{\text{old}}$ denote the final quantity of arc i at position x and the final demand at node v_i in the previous time interval, respectively.

4. NETWORK DYNAMICS WITH NONLINEAR DAMPING

In this section, we focus on the dynamics in the network on all arcs i given by constraint (1b) of the optimal control problem, i.e. the shape of the functions $f^{(i)}$ and $g^{(i)}$ governed by

$$z_t^{(i)}(x,t) + f^{(i)}(z^{(i)}(x,t),t)_x + g^{(i)}(z^{(i)}(x,t),t) = 0, \tag{2}$$

where $f^{(i)}$ denotes the flux function and $g^{(i)}$ is the damping function. Here, we restrict to functions $f^{(i)}$ that are linear in $z^{(i)}$ but nonlinear in the time and potentially nonlinear functions $g^{(i)}$ in $z^{(i)}$ and t. The damping reflects a loss in the transported quantity over time, which may be due to some physical property as for instance friction or electrical resistance. The linearity of the flux-functions will play a very important role because characteristic curves do not cross and hence no discontinuities may appear, see LeVeque (2002) for an overview. For simplicity, all investigations are executed in the 1-1-network case and can be generalized to arbitrary serial networks in a straightforward way.

We restrict to nonlinear damping functions in which the nonlinearities of time and quantity are separated, i.e.

$$g^{(i)}(z,t) = \mu_i(t)\hat{g}_i(z), \tag{3}$$

where $\mu_i \in L^1([t_0,T])$ is chosen to be a non-negative function and $\hat{g}_i$ is a Lipschitz-continuous function whose antiderivative is explicitly known. Additionally, we assume $\hat{g}_i \geq 0$ and $\hat{g}_i = 0$ only on a Lebesgue null-set. This allows for the natural choice of $\hat{g}_i(0) = 0$, but still preserves some important properties we will exploit. The flux function on arc i is chosen as $f^{(i)}(z,t) = \lambda_i(t)z$ with strictly positive velocity $\lambda_i(\cdot)$.

Next, we calculate the optimal inflow for a 1-1 network. Denote by $\mu_i(t)$ a time-dependent damping factor for arc i and $\hat{g}_i(z) = \frac{1}{\bar{g}_i(z)}$ the nonlinear damping function for arc i such that the overall damping term is given by $g_i(z,t) = \frac{\mu_i(t)}{\bar{g}_i(z)}$. We make use of the network from Figure 1, where we have one source node v_0, an inner intersection v_1 and a demand node v_2.

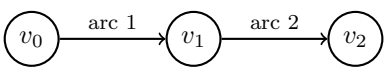


Fig. 1. The supply system as a 1-1-network with one source v_0 and one demand node v_2.

To relate the injection time at the source node v_0 and the output time at the demand node v_2 we consider a characteristic curve $(x(t),t)$ of a unit for which t_2 is the time in which the unit injected at t_{in} reaches the demand node v_2 and t_1 the time when it reaches node v_1. If the velocity functions λ_i have an antiderivative $\Lambda_i, i = 1,2$

which is invertible, for some injection time t_{in} the values of t_1 and t_2 are explicitly given by

$$\begin{aligned} t_1 &= \Lambda_1^{-1}(1+\Lambda_1(t_{\text{in}})) \\ t_2 &= \Lambda_2^{-1}(1+\Lambda_2(\Lambda_1^{-1}(1+\Lambda_1(t_{\text{in}})))), \end{aligned} \tag{4}$$

where we for simplicity assume that every arc has length 1 (see Göttlich and Schillinger (2021)).

Staying on such a characteristic curve, in a setting without damping, the transported quantity $z(x(t),t)$ remains constant. If we introduce the damping term, the quantity z is reduced by the damping. This can be formulated by the following ordinary differential equation (ODE)

$$\frac{d}{dt} z^{(i)}(x(t),t) = -g_i(z^{(i)}(x(t),t),t) = -\frac{\mu_i(t)}{\tilde{g}_i(z^{(i)}(x(t),t))}. \tag{5}$$

This ODE can be uniquely solved due the Lipschitz-continuity of $\tilde{g}$. The technique of separation of variables even allows for the calculation of an explicit solution. To do so, we assume that an antiderivative of $\tilde{g}_i$ is explicitly given by $\tilde{G}_i$ and we consider time points $t_{k-1} < t_k$. Integrating (5) both sides of the equation over t_{k-1} to t_k and applying the fundamental theorem of calculus leads to

$$\begin{aligned} &\int_{t_{k-1}}^{t_k} \frac{d}{ds}\left(z^{(i)}(x(s),s)\right) \cdot \tilde{g}_i(z^{(i)}(x(s),s))ds \\ &= -\int_{t_{k-1}}^{t_k} \mu_i(s)ds \;\Leftrightarrow\; z^{(i)}(x(t_{k-1}),t_{k-1}) \\ &= \tilde{G}_i^{-1}\left(\tilde{G}_i\left(z^{(i)}(x(t_k),t_k)\right) + \int_{t_{k-1}}^{t_k} \mu_i(s)ds\right). \end{aligned}$$

The antiderivative of $\tilde{g}_i$ is continuous and strictly increasing because $\tilde{g}_i$ is strictly positive, apart from a Lebesgue null-set. Then, there exists the inverse function $(\tilde{G}_i)^{-1}$ of $\tilde{G}_i$ and the solution is well-defined. The integral of the damping function stays bounded since $\mu_i \in L^1([t_0,T])$. Note that we are going to perform a backward calculation for the optimal inflow under the assumption that we know the corresponding optimal outflow of the network. Therefore, we consider an end value problem here, this means we have upstream information at time t_k and calculate the corresponding initial values at time t_{k-1}. We start this procedure at demand node v_2.

$$z^{(2)}(1,t_2) = \frac{f^{(2)}(z^{(2)}(1,t_2),t_2)}{\lambda_2(t_2)} \tag{6}$$

Following the characteristic curve and using the result of the ODE solution we can calculate the resulting quantity at the beginning of arc 2 at time t_1 by

$$z^{(2)}(0,t_1) = \tilde{G}_2^{-1}\left(\tilde{G}_2\left(z^{(2)}(1,t_2)\right) + \int_{t_1}^{t_2} \mu_2(s)ds\right).$$

Using this and (6), we can calculate the ingoing flux into arc 2 at time t_1 by

$$\begin{aligned} &f^{(2)}(z^{(2)}(0,t_1),t_1) = \lambda_2(t_1) z^{(2)}(0,t_1) \\ &= \lambda_2(t_1)\tilde{G}_2^{-1}\left(\tilde{G}_2\left(\frac{f^{(2)}(z^{(2)}(1,t_2),t_2)}{\lambda_2(t_2)}\right) + \int_{t_1}^{t_2} \mu_2(s)ds\right). \end{aligned}$$

Accounting for flux conservation at the intersection it must hold that

$$f^{(2)}(z^{(2)}(0,t_1),t_1) = f^{(1)}(z^{(1)}(1,t_1),t_1).$$

Then, we can deduce the corresponding quantity at the end of arc 1 by

$$\begin{aligned} z^{(1)}(1,t_1) = \frac{\lambda_2(t_1)}{\lambda_1(t_1)} \cdot \tilde{G}_2^{-1}\Bigg(&\tilde{G}_2\left(\frac{f^{(2)}(z^{(2)}(1,t_2),t_2)}{\lambda_2(t_2)}\right) \\ &+ \int_{t_1}^{t_2} \mu_2(s)ds\Bigg). \end{aligned}$$

Now again applying the ODE solution on arc 1 we get for the initial quantity

$$\begin{aligned} &z^{(1)}(0,t_{\text{in}}) \\ &= \tilde{G}_1^{-1}\Bigg(\tilde{G}_1\left[\frac{\lambda_2(t_1)}{\lambda_1(t_1)} \cdot \tilde{G}_2^{-1}\Bigg(\tilde{G}_2\left(\frac{f^{(2)}(z^{(2)}(1,t_2),t_2)}{\lambda_2(t_2)}\right) \\ &\quad + \int_{t_1}^{t_2} \mu_2(s)ds\Bigg)\right] + \int_{t_{\text{in}}}^{t_1} \mu_1(s)ds\Bigg). \end{aligned}$$

The ingoing flux into the 1-1 network is now a direct consequence of the quantity at the beginning of arc 1 and given by

$$\begin{aligned} &f^{(1)}(z^{(1)}(0,t_{\text{in}}),t_{\text{in}}) \\ &= \lambda_1(t_{\text{in}})\tilde{G}_1^{-1}\Bigg(\tilde{G}_1\left[\frac{\lambda_2(t_1)}{\lambda_1(t_1)}\tilde{G}_2^{-1}\Big(\tilde{G}_2\left(\frac{f^{(2)}(z^{(2)}(1,t_2),t_2)}{\lambda_2(t_2)}\right) \\ &\quad + \int_{t_1}^{t_2} \mu_2(s)ds\Big)\right] + \int_{t_{\text{in}}}^{t_1} \mu_1(s)ds\Bigg). \end{aligned}$$

Since the optimal outflow of the network should match the corresponding conditional expected demand we obtain for the optimal inflow

$$\begin{aligned} &u(t_{\text{in}}) \\ &= \lambda_1(t_{\text{in}}) \cdot \tilde{G}_1^{-1}\Bigg(\tilde{G}_1\left[\frac{\lambda_2(t_1)}{\lambda_1(t_1)} \cdot \tilde{G}_2^{-1}\Bigg(\tilde{G}_2\left(\frac{\mathbb{E}\left[D_{t_2}|\mathcal{F}_{\hat{t}}\right]}{\lambda_2(t_2)}\right) \\ &\quad + \int_{t_1}^{t_2} \mu_2(s)ds\Bigg)\right] + \int_{t_{\text{in}}}^{t_1} \mu_1(s)ds\Bigg), \end{aligned}$$

where $\hat{t} < t_{\text{in}}$ is the time of the latest demand update. Iteratively, this procedure can be extended to larger 1-1 networks.

5. NUMERICAL EXPERIMENTS

Finally, we numerically compare the solutions to (1) for different nonlinear damping functions $\hat{g}_i(z)$. Classical choices for such nonlinear damping functions are monomials (Ikeda et al. (2017)), i.e. functions of the type

$$\hat{g}^{(n)}(z) = C_n z^n,$$

where C_n is a constant depending on the degree which has to be determined. By construction it holds that $\hat{g}^{(n)}(0) = 0$. For better illustration purposes, we choose the same damping functions on both arcs. To be able to compare the damping functions we request that

$$\int_0^{\frac{1}{10}} \hat{g}^{(n)}(z)dz = \frac{1}{20}, \quad n \in \mathbb{N}$$

and choose the constant C_n accordingly. For our analysis we compare monomials up to order 4, which are given by

$$\begin{aligned} &\hat{g}^{(1)}(z) = z, \quad \hat{g}^{(2)}(z) = 15z^2, \\ &\hat{g}^{(3)}(z) = 200z^3, \quad \hat{g}^{(4)}(z) = 2500z^4. \end{aligned} \tag{7}$$

Additionally we consider scenarios without any damping, i.e $\mu_i(t) = 0$. Note that also different approaches for $\hat{g}$ can

be used, but since monomials give a good impression on the different scaling we do not extend the study at that point.

For the demand process we choose a Jacobi process with $\theta(t) = 0.45 + 0.2\sin(\pi t + 1)$, $\kappa = 2$, $\sigma = 0.9$ and initial demand of $D_0 = 0.4$.

A truncated Euler-Maruyama scheme is used for the simulation of the Jacobi process. We consider a time grid $(t_j)_{j\in\mathbb{N}}$ with temporal stepsize $\Delta t = \frac{1}{1000}$, i.e.

$$D^*_{j+1} = D_j + \Delta t\kappa(\theta(t_j) - D_j) + \sigma\sqrt{\Delta t D_j(1-D_j)}X_j,$$

where X_j is a realization of a standard normal distributed random variable. To numerically avoid values outside the interval $[0,1]$ for D_{j+1} due to the unboundedness of X_j, we add a truncation into the Euler-Maruyama scheme such that the process D^*_{j+1} is reflected into $[0,1]$:

$$D_{j+1} = \begin{cases} 1, & D^*_{j+1} \geq 1 \\ D^*_{j+1}, & D^*_{j+1} \in (0,1) \\ 0, & D^*_{j+1} \leq 0. \end{cases}$$

The time horizon is chosen to $T = 2.5$ and we consider an update strategy at the source node with 7 equidistant updates. The flux functions on the arcs are all linear with respect to the quantity $z^{(i)}$ and have the time-dependent factor $\lambda_i(t) = 14 + \sin(2\pi t)$. The temporal factor of the damping in (3) is chosen to be $\mu_i(t) = 5 + \sin(\pi t)$, where $\hat{g}$ are the monomials from (7).

For the discretization of (2) we use an adaptive upwind-scheme with spatial stepsize $\Delta x = \frac{1}{200}$ on a grid $(x_l)_{l\in\mathbb{N}}$. The damping is incorporated by a splitting algorithm which first performs the step from the upwind-scheme and then applies the damping in a second step. The temporal step sizes are chosen such that the CFL-condition (LeVeque (2002)) is satisfied with equality in every time step, i.e. $\frac{\Delta t_j^{(i)}}{\Delta x}\lambda_i(t_j^{(i)}) = 1$. Therefore, the temporal grids depend on the velocity functions of the particular arc:

$$\tilde{z}_l^{(i),j+1} = z_l^{(i),j} + \frac{\Delta t_j^{(i)}}{\Delta x}\lambda_l(t_j^{(i)})\left(z_l^{(i),j} - z_{l-1}^{(i),j}\right).$$

In a second step we take into account the damping and calculate $z_l^{(i),j+1}$ by

$$z_l^{(i),j+1} = \tilde{z}_l^{(i),j+1} - \Delta t_j^{(i)} g_i\left(\tilde{z}_l^{(i),j+1}, t_{j+1}^{(i)}\right).$$

In Figure 2 we present results for a 1-1 network and show the inflows for five different types of damping (from no damping to a quartic damping monomial in (7)) in the upper part as well as the outflow and demand at the demand node in the lower part. Note that the supplies and demands do not differ significantly among the different damping settings. The shape of the curves is mainly influenced by the temporal nonlinearities in the flux functions and the sinusoidal rhythm of the demand mean reversion levels. First, we observe that in a setting without damping the inflow is always below the inflows of the damped setting. As a second observation, we see that the inflows for the monomials are ordered differently depending on the amount of the inflow and the corresponding quantities. For larger quantities the inflows for the second and third order monomial damping function are higher, whereas for lower inflows the inflows resulting from linear damping have the highest values. In the outflow plot we observe that the supply adjust overall to the shape of demand, where the jumps in supply result from demand updates.

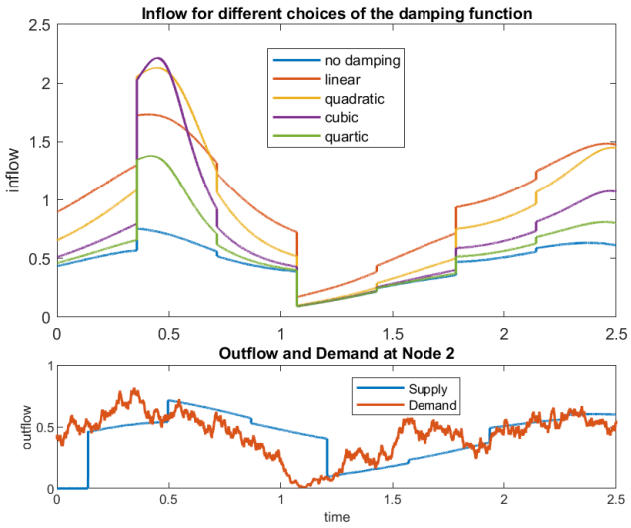


Fig. 2. Comparison of different inflows for different choices of the damping function and the corresponding supply and demand for one realization.

REFERENCES

Banda, M., Herty, M., and Klar, A. (2006). Gas flow in pipeline networks. *Netw. Heterog. Media*, 1(1), 41–56.

Bressan, A., Čanić, S., Garavello, M., Herty, M., and Piccoli, B. (2014). Flows on networks: recent results and perspectives. *EMS Surv. Math. Sci.*, 1(1), 47–111.

Coskun, S. and Korn, R. (2021). *Modeling the Intraday Electricity Demand in Germany*, 3–23. Springer International Publishing, Cham.

D'Apice, C., Göttlich, S., Herty, M., and Piccoli, B. (2010). *Modeling, Simulation, and Optimization of Supply Chains.* Society for Industrial and Applied Mathematics.

Delbaen, F. and Shirakawa, H. (2002). An interest rate model with upper and lower bounds. *Asia-Pacific Financial Markets*, 9, 191–209.

Göttlich, S., Herty, M., and Schillen, P. (2015). Electric transmission lines: Control and numerical discretization. *Optimal Control Applications and Methods*, 980– 995.

Göttlich, S., Korn, R., and Lux, K. (2019). Optimal control of electricity input given an uncertain demand. *Mathematical Methods of Operations Research*, 90, 1–28.

Göttlich, S. and Schillinger, T. (2021). Control strategies for transport networks under demand uncertainty.

Ikeda, M., Inui, T., and Yuta, W. (2017). The cauchy problem for the nonlinear damped wave equation with slowly decaying data. *Nonlinear Differential Equations and Applications NoDEA*, 24.

Klenke, A. (2020). *Probability Theory : A Comprehensive Course.* Universitext. 3rd ed. 2020. edition.

LeVeque, R.J. (2002). *Finite Volume Methods for Hyperbolic Problems.* Cambridge Texts in Applied Mathematics. Cambridge University Press.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Local turnpike analysis for discrete time discounted optimal control

Lars Grüne* **and Lisa Krügel***

* *Chair of Applied Mathematics, Mathematical Institute, Universität Bayreuth, Germany, (e-mail: lars.gruene, lisa.kruegel@uni-bayreuth.de)*

Keywords: Discounted Optimal Control, Dissipativity, Turnpike

1. INTRODUCTION

Recent results in the literature have provided connections between the turnpike property, near optimality of closed-loop solutions, and strict dissipativity. In this talk, based on the recent paper Grüne and Krügel (2021) (to which we refer for all proofs), we consider optimal control problems with discounted stage cost. In contrast to non-discounted optimal control problems, it is more likely that several asymptotically stable optimal equilibria coexist. Due to the discounting and transition cost from a local to the global equilibrium, it may be more favourable staying in a local equilibrium than moving to the global "cheaper" equilibrium. In this talk, we propose a local notion of discounted strict dissipativity and a local turnpike property, both depending on the discount factor. Using these concepts, we investigate the local behaviour of (near-)optimal trajectories and develop conditions on the discount factor to ensure convergence to a local asymptotically stable optimal equilibrium.

2. SETTING

We consider discrete time nonlinear systems of the form

$$x(k+1) = f(x(k), u(k)), \quad x(0) = x_0 \tag{1}$$

for a map $f : X \times U \to X$, where X and U are normed spaces. We impose the constraints $(x, u) \in \mathbb{Y} \subset X \times U$ on the state x and the input u and define $\mathbb{X} := \{x \in X \mid \exists u \in U : (x, u) \in \mathbb{Y}\}$ and $\mathbb{U} := \{u \in U \mid \exists x \in X : (x, u) \in \mathbb{Y}\}$. A control sequence $u \in \mathbb{U}^N$ is called admissible for $x_0 \in \mathbb{X}$ if $(x(k, x_0), u(k)) \in \mathbb{Y}$ for $k = 0, \dots, N-1$ and $x(N) \in \mathbb{X}$. In this case, the corresponding trajectory $x(k)$ is also called admissible. The set of admissible control sequences is denoted by $\mathbb{U}^N(x_0)$. Likewise, we define $\mathbb{U}^\infty(x_0)$ as the set of all control sequences $u \in \mathbb{U}^\infty$ with $(x(k, x_0), u(k)) \in \mathbb{Y}$ for all $k \in \mathbb{N}_0$. Furthermore, we assume that $\mathbb{X}$ is controlled invariant, i.e. that $\mathbb{U}^\infty(x_0) \neq \emptyset$ for all $x_0 \in \mathbb{X}$. The trajectories of (1) are denoted by $x_u(k, x_0)$ or simply by $x(k)$ if there is no ambiguity about x_0 and u.

We will make use of comparison-functions defined by

$$\mathcal{K} := \{\alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+ | \alpha \text{ is continuous and strictly increasing with } \alpha(0) = 0\}$$
$$\mathcal{K}_\infty := \{\alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+ | \alpha \in \mathcal{K}, \alpha \text{ is unbounded}\}$$

⋆ The authors are supported by DFG Grant Gr 1569/13-2.

Moreover, with $\mathcal{B}_\varepsilon(x_0)$ we denote the open ball with radius $\varepsilon > 0$ around x_0.

In this talk we consider infinite horizon discounted optimal control problems, i.e. problems of the type

$$\min_{u \in \mathbb{U}^\infty(x_0)} J_\infty(x_0, u) \quad \text{with } J_\infty(x_0, u) = \sum_{k=0}^{\infty} \beta^k \ell(x(k, x_0), u(k)). \tag{2}$$

Herein, the number $\beta \in (0, 1)$ is called the discount factor. The optimal value function of the problem is defined by

$$V_\infty(x_0) := \min_{u \in \mathbb{U}^\infty(x_0)} J_\infty(x_0, u).$$

3. GLOBAL TURNPIKE

The central structural assumption on this optimal control problem is a strict dissipativity condition. Here we first introduce a global version, which will later be relaxed to a local assumption:

We say that the system (1) is discounted strictly dissipative at an equilibrium (x^β, u^β) if there exists a storage function $\lambda : \mathbb{X} \to \mathbb{R}$ bounded from below with $\lambda(x^\beta) = 0$ and a class $\mathcal{K}_\infty$-function α such that the inequality

$$\ell(x, u) - \ell(x^\beta, u^\beta) + \lambda(x) - \beta\lambda(f(x, u)) \geq \alpha(\|x - x^\beta\|) \tag{3}$$

holds for all $(x, u) \in \mathbb{Y}$ with $f(x, u) \in \mathbb{X}$. We remark that this discounted version of discrete-time strict dissipativity was studied in Grüne et al. (2021), based on the original undiscounted definitions from Byrnes and Lin (1994) and Willems (1972).

For a strictly dissipative problem we can define the modified or rotated stage cost

$$\widetilde{\ell}(x, u) = \ell(x, u) - \ell(x^\beta, u^\beta) + \lambda(x) - \beta\lambda(f(x, u)). \tag{4}$$

It was shown in Grüne et al. (2021) that the optimal trajectories of the optimal control problems with stage cost ℓ and $\widetilde{\ell}$ coincide if the storage function λ is bounded.

With a combination of arguments from Gaitsgory et al. (2018), one can prove that if the optimal value function $\widetilde{V}_\infty$ of the modified problem satisfies

$$\widetilde{V}_\infty(x) \leq \alpha_V(\|x - x^\beta\|) \quad \text{and} \quad \widetilde{V}_\infty(x) \leq C \inf_{u \in \mathbb{U}} \tilde{\ell}(x, u) \tag{5}$$

for all $x \in \mathbb{X}$, a function $\alpha_V \in \mathcal{K}_\infty$, and a constant $C \geq 1$ exist satisfying

$$C < 1/(1-\beta). \tag{6}$$

Then the optimal control problem has the following turnpike property, cf. (Grüne et al., 2017, Definition 4.2):

For each $\varepsilon > 0$ and each bounded set $\mathbb{X}_b \subset \mathbb{X}$ there exists a constant $P > 0$ such that for each $M \in \mathbb{N}$ there is a $\delta > 0$, such that for all $x_0 \in \mathbb{X}_b$ and $u \in \mathbb{U}^\infty(x_0)$ with $J_\infty(x_0, u) \leq V_\infty(x_0) + \delta$, the set $\mathcal{Q}(x_0, u, \varepsilon, M, \beta) := \{k \in \{0, \dots, M\} \mid \|x_u(k, x_0) - x^\beta\| \geq \varepsilon\}$ has at most P elements.

Figure 1 gives an illustration of the set $\mathcal{Q}(x_0, u, \varepsilon, M, \beta)$ from this definition.

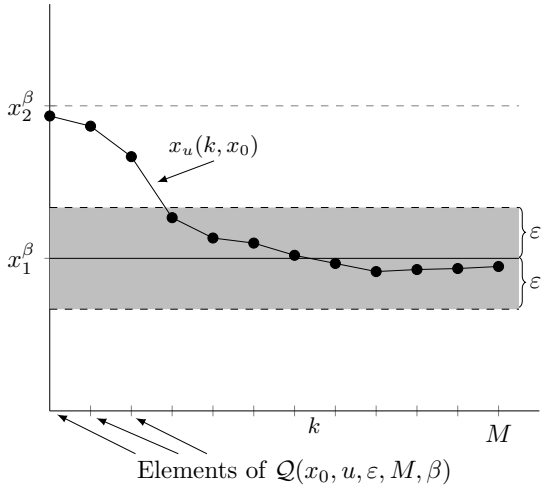


Fig. 1. Illustration of the set $\mathcal{Q}(x_0, u, \varepsilon, M, \beta)$

4. LOCAL TURNPIKE

In discounted optimal control, several optimal equilibria (x_l^β, u_l^β) with different optimal values $\ell(x_l^\beta, u_l^\beta)$ can coexist, even if the system is completely controllable. This is because the transition cost for passing from an equilibrium to a region in the state space with lower cost may dominate the benefit from reaching this "cheaper" region, because the discounting reduces the relative weight of this benefit. This means that local turnpike properties may emerge, in which the solutions only approach a certain equilibrium x_l^β if the initial value is close to this equilibrium, and show a different long time behavior otherwise. The closer the discount rate is to 1, the less weight is put on the transition cost, hence the less likely a local turnpike behavior becomes if the equilibrium is not globally strictly dissipative. Hence, a local turnpike property is more likely to occur for discount rates $\beta \ll 1$ than for discount rates $\beta \approx 1$. On the other hand, the conditions (5) and (6) will also be needed for a local turnpike property and thus provide a lower bound on the possible discount rates. Hence, we can expect that there is an interval of discount rates for which a local turnpike property occurs.

In order to formalize this intuitive description, we first introduce a local strict dissipativity property:

Given a discount factor $\beta \in (0,1)$, we say that the system (1) is locally discounted strictly dissipative at an equilibrium (x_l^β, u_l^β) if there exists a storage function $\lambda : \mathbb{X} \to \mathbb{R}$ bounded from below with $\lambda(x_l^\beta) = 0$ and a class $\mathcal{K}_\infty$-function α_β such that the inequality

$$\ell(x,u) - \ell(x_l^\beta, u_l^\beta) + \lambda(x) - \beta\lambda(f(x,u)) \geq \alpha_\beta(\|x - x_l^\beta\|) \tag{7}$$

holds for all $(x,u) \in \mathbb{X}_\mathcal{N} \times \mathbb{U}$ with $f(x,u) \in \mathbb{X}$.

Further, we say that system (1) is locally discounted strictly (x,u)-dissipative at the equilibrium (x_l^β, u_l^β) with supply rate $s : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ if the same holds with the inequality

$$\begin{aligned} &\ell(x,u) - \ell(x_l^\beta, u_l^\beta) \\ &+\lambda(x) - \beta\lambda(f(x,u)) \geq \alpha_\beta(\|(x - x_l^\beta\| + \|u - u_l^\beta)\|). \end{aligned} \tag{8}$$

The main theorem we are going to present in this talk is now the following. For a proof see (Grüne and Krügel, 2021, Theorem 6.1).

Theorem 1. Consider a discounted optimal control problem (2) subject to system (1) with f continuous and stage cost ℓ bounded from below. Assume local strict (x,u)-dissipativity at an equilibrium (x_l^β, u_l^β) according to (8) with bounded storage function λ. Assume furthermore that there is an interval $(\beta_1, \beta^\star)$ of discount rates, such that for each $\beta \in (\beta_1, \beta^\star)$ the inequalities (5) and (6) hold for all $x \in \mathbb{X}_\mathcal{N}$ with x_l^β in place of x^β.

Then there is $\beta_2 \in (0,1)$, depending on the problem data, such that for all $\beta \in (\beta_1, \beta_2)$ there exists a neighbourhood $\mathcal{N}$ of x_l^β on which the system exhibits a local turnpike property in the following sense:

For each $\varepsilon > 0$ there exist a constant $P > 0$ such that for each $M \in \mathbb{N}$ there is a $\delta > 0$, such that for all $x_0 \in \mathcal{N}$ and all $u \in \mathbb{U}^\infty(x_0)$ with $J_\infty(x_0, u) \leq V_\infty(x_0) + \delta$, the set $\mathcal{Q}(x, u, \varepsilon, M, \beta) := \{k \in \{0, \dots, M\} \mid \|x_u(k, x_0) - x_l^\beta\| \geq \varepsilon\}$ has at most P elements.

Particularly, if $J_\infty(x_0, u) = V_\infty(x_0)$, i.e., if the trajectory is optimal, then for each $\varepsilon > 0$ the set $\mathcal{Q}(x, u, \varepsilon, \infty, \beta) := \bigcup_{M \in \mathbb{N}} \mathcal{Q}(x, u, \varepsilon, M, \beta)$ has at most P elements, implying the convergence $x_u(k, x_0) \to x_l^\beta$ as $k \to \infty$.

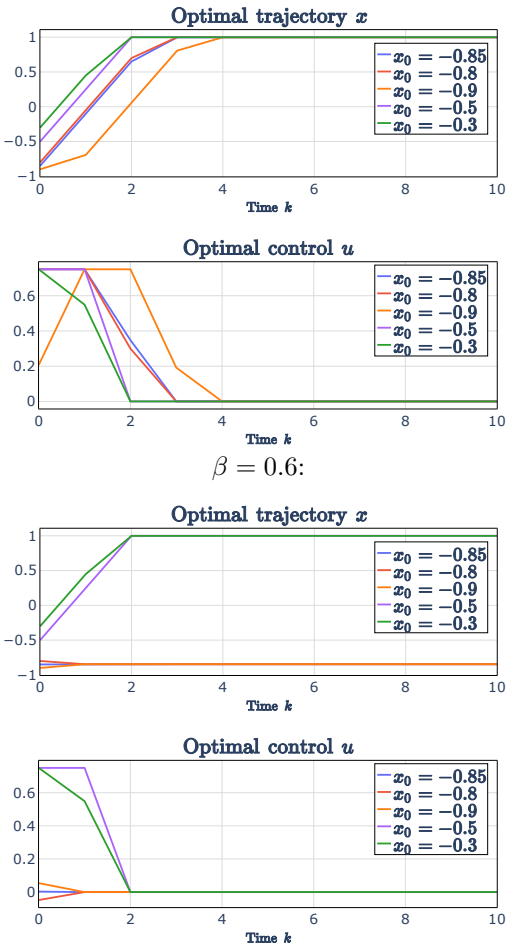
We note that the fact that β_2 depends on the problem data is important for this theorem to yield a meaningful statement, because otherwise the assertion is always trivially true for any $\beta_2 < \beta_1$ since then the interval (β_1, β_2) is empty. We will explain in the talk that β_2 depends on the the cost to leave the neighborhood $\mathbb{X}_\mathcal{N}$ and the minimum of $\widetilde{\ell}$ over $\mathbb{Y}$. The larger these two values become, the larger β_2 becomes.

5. ILLUSTRATIVE EXAMPLE

For illustrating our results, consider the one-dimensional dynamics $f(x,u) = x + u$ and the stage cost $\ell(x,u) = x^4 - \frac{1}{4}x^3 - \frac{7}{4}x^2$. It turns out that the resulting optimal control problem has two optimal equilibria in $x_l^\beta = \frac{3-\sqrt{905}}{32} \approx -0.846$ and $x_g^\beta = \frac{3+\sqrt{905}}{32} \approx 1.034$, where $\ell(x_g^\beta) < \ell(x_l^\beta)$. Note that the equilibria do not depend on β in this example, while in general they may do so. Both optimal equilibria are strictly dissipative, x_l^β only locally and x_g^β also globally. Using the construction in the proof of (Grüne and Krügel, 2021, Theorem 6.1), one computes that $\beta_2 \approx 0.67$. Hence one would expect that a local turnpike property is visible for β below this value and

ceases to exist for β above this value. The numerical simulations of optimal trajectories in Figure 5 show that this is precisely what happens. The figure shows optimal trajectories and control functions with initial values x_0 between -1 and -0.3 and $\beta = 0.7$ and $\beta = 0.6$. While for $\beta = 0.7$ all solutions converge to the globally optimal equilibrium x_g^β, for $\beta = 0.6$ solutions with x_0 sufficiently close to x_l^β converge to x_l^β. Hence, this illustrates the existence of the local turnpike property.

$\beta = 0.7$:



$\beta = 0.6$:

Fig. 2. Optimal trajectories and control functions for $\beta = 0.7$ (above) and $\beta = 0.6$ (below) for different initial values x_0

REFERENCES

Byrnes, C.I. and Lin, W. (1994). Losslessness, feedback equivalence, and the global stabilization of discrete-time nonlinear systems. *IEEE Trans. Automat. Control*, 39(1), 83–98.

Gaitsgory, V., Grüne, L., Höger, M., Kellett, C.M., and Weller, S.R. (2018). Stabilization of strictly dissipative discrete time systems with discounted optimal control. *Automatica*, 93, 311–320. doi: 10.1016/j.automatica.2018.03.076.

Grüne, L. and Krügel, L. (2021). Local turnpike analysis using local dissipativity for discrete time discounted optimal control. *Appl. Math. Optim.*, 84, 1585–1606. doi:https://doi.org/10.1007/s00245-021-09805-4.

Grüne, L., Müller, M.A., Kellett, C.M., and Weller, S.R. (2021). Strict dissipativity for discrete time discounted optimal control problems. *Math. Control Relat. Fields*, 11, 771–796.

Grüne, L., Kellett, C., and Weller, S. (2017). On the relation between turnpike properties for finite and infinite horizon optimal control problems. *Journal of Optimization Theory and Applications*, 173. doi:10.1007/s10957-017-1103-6.

Willems, J.C. (1972). Dissipative dynamical systems. I. General theory. *Arch. Rational Mech. Anal.*, 45, 321–351.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Efficient algorithms for certain high-dimensional optimal control problems based on novel Hopf-type and Lax-Oleinik-type representation formulas

Paula Chen [*] **, Jérôme Darbon** [**] **, Tingwei Meng** [***]

[*] *Division of Applied Mathematics, Brown University, Providence, RI, USA (e-mail: paula_chen@brown.edu)*
[**] *Division of Applied Mathematics, Brown University, Providence, RI, USA (e-mail: jerome_darbon@brown.edu)*
[***] *Department of mathematics, UCLA, Los Angeles, CA, USA (e-mail: tingwei@math.ucla.edu)*

Abstract: Solving high-dimensional optimal control problems and their corresponding Hamilton-Jacobi partial differential equations is an important but challenging problem. In particular, handling optimal control problems with state-dependent running costs or constraints on the control presents an additional challenge. We present two representation formulas: one is a Hopf-type representation formula for solving a class of optimal control problems with certain non-smooth state-dependent running costs, and the other is a Lax-Oleinik-type representation formula for solving a class of optimal control problems with certain control constraints. Based on these formulas, we propose efficient algorithms that overcome the curse of dimensionality. As such, our proposed methods have the potential to serve as a building block for solving more complicated high-dimensional optimal control problems in real-time.

Keywords: Optimal Control, Hamilton-Jacobi Partial Differential Equations, Grid-Free Numerical Methods, High Dimensions

1. INTRODUCTION

It is well-studied that solving optimal control problems is related to solving Hamilton-Jacobi partial differential equations (HJ PDEs). However, the computational complexity of standard grid-based numerical algorithms for solving HJ PDEs scales exponentially with respect to the dimension. This exponential scaling in dimension is often referred to as the "curse of dimensionality" (CoD) [Bellman (1961)]. Due to the CoD, these grid-based methods are infeasible for solving high-dimensional problems (e.g., dimensions greater than five) and thus are infeasible for many practical optimal control applications, which often have dimension much greater than five.

To overcome the CoD, many grid-free methods approximate the original optimal control problem by some simpler, more easily computable optimal control problems. Thus, an important research direction is to enlarge the class of optimal control problems with easily computable solutions; these problems and their corresponding exact solvers can then serve as building blocks for solving more complicated optimal control problems. Some well-known techniques for solving optimal control problems that often serve as these building blocks include: the linear-quadratic regulator [Li and Todorov (2004); Sideris and Bobrow (2005); McEneaney (2006); Coupechoux et al. (2019)], the Hopf and Lax-Oleinik representation formulas [Darbon (2015); Darbon and Meng (2020); Darbon and Osher (2016); Yegorov and Dower (2017)], and the max-plus (or min-plus) technique [Akian et al. (2006, 2008); Dower et al. (2015); Fleming and McEneaney (2000); Gaubert et al. (2011); McEneaney (2006, 2007); McEneaney et al. (2008); McEneaney and Kluberg (2009)]. However, there are still many more classes of optimal control problems that cannot be solved (exactly) using these techniques. For example, optimal control problems with state-dependent running costs or constraints on the control are, in general, difficult to solve without approximations.

In this extended abstract, we summarize our recent works [Chen et al. (2021a,b)], which provide analytical solutions to two classes of high-dimensional optimal control problems. The first class involves certain state-dependent non-smooth running costs and is summarized in Section 2. The second class involves certain state-dependent running costs and control constraints and is summarized in Section 3. For both classes of problems, we also provide efficient numerical solvers that do not rely on approximations of the corresponding optimal control problem and HJ PDE and are numerically shown to overcome the CoD.

[⋆] This research is supported by NSF 1820821 and AFOSR MURI FA9550-20-1-0358. P.C. is supported by the SMART Scholarship, which is funded by USD/R&E (The Under Secretary of Defense-Research and Engineering), National Defense Education Program (NDEP) / BA-1, Basic Research.

2. THE FIRST CLASS OF OPTIMAL CONTROL PROBLEMS

We consider the following high-dimensional optimal control problem with state-dependent non-smooth running cost:

$$V(\boldsymbol{x},t)=\inf\left\{\int_0^t\left(\frac{1}{2}\|\dot{\boldsymbol{x}}(s)\|^2-U(\boldsymbol{x}(s))\right)ds+\Phi(\boldsymbol{x}(0))\colon \boldsymbol{x}(t)=\boldsymbol{x}\right\}, \tag{1}$$

where $\|\cdot\|$ denotes the ℓ^2-norm, $\boldsymbol{x}\in\mathbb{R}^n$ and $t>0$ are the terminal position and time horizon, respectively, the initial cost $\Phi\colon\mathbb{R}^n\to\mathbb{R}$ is a convex function, and the function $U\colon\mathbb{R}^n\to(-\infty,0]$ satisfies $U(\boldsymbol{x})=\sum_{i=1}^n U_i(x_i)$ for any $\boldsymbol{x}=(x_1,\dots,x_n)\in\mathbb{R}^n$, where each function $U_i\colon\mathbb{R}\to(-\infty,0]$ is the 1-homogeneous concave function defined by

$$U_i(x)=\begin{cases}-a_i x & x\geq 0,\\ b_i x & x<0,\end{cases}$$

with positive constants a_i and b_i. In the remainder of this abstract, we use bold characters to denote high-dimensional vectors in $\mathbb{R}^n$, and we use x_i to denote the i-th component of a high-dimensional vector $\boldsymbol{x}\in\mathbb{R}^n$. The corresponding HJ PDE reads:

$$\begin{cases}\dfrac{\partial V}{\partial t}(\boldsymbol{x},t)+\dfrac{1}{2}\|\nabla_{\boldsymbol{x}}V(\boldsymbol{x},t)\|^2+U(\boldsymbol{x})=0 & \boldsymbol{x}\in\mathbb{R}^n, t\in(0,+\infty),\\ V(\boldsymbol{x},0)=\Phi(\boldsymbol{x}) & \boldsymbol{x}\in\mathbb{R}^n,\end{cases} \tag{2}$$

where the initial condition Φ and the potential energy U are the corresponding functions in (1).

In Chen et al. (2021a), we show that under some assumptions, the viscosity solution $V\colon\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}$ is defined by the following Hopf-type formula:

$$V(\boldsymbol{x},t):=\sup_{\boldsymbol{p}\in\mathbb{R}^n}\left\{\sum_{i=1}^n V(x_i,t;p_i,a_i,b_i)-\Phi^*(\boldsymbol{p})\right\}, \tag{3}$$

for any $\boldsymbol{x}\in\mathbb{R}^n$, $t\geq 0$, where the function $(x_i,t)\mapsto V(x_i,t;p_i,a_i,b_i)$ on the right-hand side is a continuously differentiable function defined explicitly in Chen et al. (2021a). Moreover, for a positive time horizon $t>0$, the function value $V(\boldsymbol{x},t)$ defined in (3) is finite, and the maximizer in (3) exists and is unique. We denote the unique maximizer by $\boldsymbol{p}^*=(p_1^*,\dots,p_n^*)\in\mathbb{R}^n$. The optimal trajectory $[0,t]\ni s\mapsto\boldsymbol{\gamma}(s;\boldsymbol{x},t)\in\mathbb{R}^n$ of the optimal control problem (1) is then given by

$$\boldsymbol{\gamma}(s;\boldsymbol{x},t):=\left(\gamma(s;x_1,t,p_1^*,a_1,b_1),\dots,\gamma(s;x_n,t,p_n^*,a_n,b_n)\right), \tag{4}$$

for any $s\in[0,t]$, where the i-th element $\gamma(s;x_i,t,p_i^*,a_i,b_i)$ on the right-hand side is a one-dimensional function whose explicit formula is given in Chen et al. (2021a). Note that the components of $\boldsymbol{\gamma}(s;\boldsymbol{x},t)$ are independent from each other, and hence, they can be computed in parallel as long as $\boldsymbol{p}^*$ is known.

Now, we will present efficient algorithms for solving the optimal control problem (1) and the corresponding HJ PDEs (2). All of the numerical examples for this class of problems were run on an 8th Gen Intel Laptop Core i5-8250U with a 1.60GHz processor. First, consider the quadratic initial condition, i.e., set

$$\Phi(\boldsymbol{x})=\frac{1}{2\lambda}\|\boldsymbol{x}-\boldsymbol{y}\|^2+\alpha\quad\forall\boldsymbol{x}\in\mathbb{R}^n,$$

where $\boldsymbol{y}\in\mathbb{R}^n$, $\lambda>0$, and $\alpha\in\mathbb{R}$ are some known parameters. The i-th component of the maximizer $\boldsymbol{p}^*$ in (3) is the proximal point of $p\mapsto-\frac{1}{\lambda}V(x_i,t;p,a_i,b_i)$ at $-\frac{y_i}{\lambda}$, which can be efficiently computed using the algorithm proposed in Chen et al. (2021a). Then, the solution to the optimal control problem (1) and the corresponding HJ PDE (2) can be computed using (4) and (3), respectively. In Chen et al. (2021a), we see that it takes less than 2×10^{-6} seconds on average to compute the solution at one point in a 16-dimensional problem with quadratic initial condition $\Phi(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{x}-\mathbf{1}\|^2$, which demonstrates the efficiency of our proposed solver in high dimensions.

Next, consider more general convex initial conditions. Note that (3) can be written as a convex optimization problem and thus can be solved using many proximal point based optimization algorithms. For illustration, we present an ADMM scheme (see Glowinski (2014); Boyd et al. (2011)), whose k-th step contains the updates of $\boldsymbol{v}^{k+1}$, $\boldsymbol{d}^{k+1}$, and $\boldsymbol{w}^{k+1}$. First, we update $\boldsymbol{v}^{k+1}\in\mathbb{R}^n$ by

$$\boldsymbol{v}^{k+1}=\operatorname*{arg\,min}_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\Phi^*(\boldsymbol{v})+\frac{\lambda}{2}\left\|\boldsymbol{v}-\boldsymbol{d}^k+\boldsymbol{w}^k\right\|^2\right\},$$

which can be efficiently computed if there exists an efficient numerical algorithm for computing the proximal point of Φ or Φ^*. Then, we update $\boldsymbol{d}^{k+1}\in\mathbb{R}^n$, where the i-th element d_i^{k+1} is computed by

$$d_i^{k+1}=\operatorname*{arg\,min}_{d_i\in\mathbb{R}}\left\{-V(x_i,t;d_i,a_i,b_i)+\frac{\lambda}{2}(v_i^{k+1}-d_i+w_i^k)^2\right\}. \tag{5}$$

Note that (5) is a one-dimensional convex optimization problem whose objective function has an explicit formula. Therefore, the minimizer d_i^{k+1} can be efficiently computed using the algorithm proposed in Chen et al. (2021a). The last step is to update $\boldsymbol{w}^{k+1}\in\mathbb{R}^n$ by

$$\boldsymbol{w}^{k+1}=\boldsymbol{w}^k+\boldsymbol{v}^{k+1}-\boldsymbol{d}^{k+1}. \tag{6}$$

We set the stopping criteria to be

$$\max\{\|\boldsymbol{v}^{N-1}-\boldsymbol{v}^N\|^2,\|\boldsymbol{d}^{N-1}-\boldsymbol{d}^N\|^2,\|\boldsymbol{v}^N-\boldsymbol{d}^N\|^2\}\leq\epsilon. \tag{7}$$

After N steps when the stopping criteria (7) is met, we set the maximizer of (3) to be $\boldsymbol{p}^N=\boldsymbol{v}^N$, and then output the optimal trajectory by

$$\hat{\boldsymbol{\gamma}}(s;\boldsymbol{x},t)=(\gamma(s;x_1,t,p_1^N,a_1,b_1),\dots,\gamma(s;x_n,t,p_n^N,a_n,b_n)), \tag{8}$$

where the i-th component $\gamma(s;x_i,t,p_i^N,a_i,b_i)$ is defined in Chen et al. (2021a). The numerical solution to the HJ PDE at $(\boldsymbol{x},t)$ is given by

$$\hat{V}(\boldsymbol{x},t)=\sum_{i=1}^n V\left(x_i,t;p_i^N,a_i,b_i\right)-\Phi^*\left(\boldsymbol{p}^N\right), \tag{9}$$

where the i-th component $V\left(x_i,t;p_i^N,a_i,b_i\right)$ in the summation is defined in Chen et al. (2021a). In Chen et al. (2021a), we present several numerical examples that demonstrate the efficiency of our proposed algorithm.

For instance, consider the initial condition given by

$$\Phi(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{x}-\mathbf{1}\|_1^2, \tag{10}$$

where $\|\cdot\|_1$ denotes the ℓ^1-norm and $\mathbf{1}$ denotes the vector in $\mathbb{R}^n$ whose elements are all one. Then, the average time per call of our algorithm to evaluate the solution of the HJ PDE (2) is summarized below for various dimensions n.

n	4	8	12	16
running time (s)	2.1192e-05	9.4819e-05	2.0531e-04	3.2751e-04

Specifically, we see that our algorithm takes less than 4×10^{-4} seconds on average to compute the solution at one point in 16 dimensions, which demonstrates its efficiency even in high dimensions.

3. THE SECOND CLASS OF OPTIMAL CONTROL PROBLEMS

We consider the following high-dimensional optimal control problem with state-dependent running cost and control constraints:

$$V(\boldsymbol{x},t) = \min\left\{ \int_0^t \frac{\|\boldsymbol{x}(s)\|^2}{2} ds + \Phi(\boldsymbol{x}(0)) \colon \boldsymbol{x}(t) = \boldsymbol{x}, \right. \\ \left. \dot{\boldsymbol{x}}(s) \in \prod_{i=1}^n [-b_i, a_i] \ \forall s \in (0,t) \right\}, \tag{11}$$

where $\boldsymbol{x}$ and t are the terminal position and time horizon, respectively, $\{a_i\}$ and $\{b_i\}$ are positive scalars which provide restrictions on the control (velocity) $\dot{\boldsymbol{x}}$, and the initial cost $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ is a lower semi-continuous function. The corresponding HJ PDE reads:

$$\begin{cases} \dfrac{\partial V}{\partial t}(\boldsymbol{x},t) + \displaystyle\sum_{i=1}^n K_i\left(\frac{\partial V(\boldsymbol{x},t)}{\partial x_i}\right) - \frac{1}{2}\|\boldsymbol{x}\|^2 = 0 & \boldsymbol{x}\in\mathbb{R}^n, t\in(0,+\infty), \\ V(\boldsymbol{x},0) = \Phi(\boldsymbol{x}) & \boldsymbol{x}\in\mathbb{R}^n, \end{cases} \tag{12}$$

where each function $K_i\colon \mathbb{R}\to\mathbb{R}$ is the 1-homogeneous convex function defined by

$$K_i(x) = \begin{cases} a_i x & x \ge 0, \\ -b_i x & x < 0, \end{cases}$$

and the initial condition is given by the initial cost Φ in (11).

In Chen et al. (2021b), we show that under some assumptions, the viscosity solution $V\colon \mathbb{R}^n\times\mathbb{R}\to\mathbb{R}$ is defined by the following Lax-Oleinik-type representation formula:

$$V(\boldsymbol{x},t) := \inf_{\boldsymbol{u}\in\prod_{i=1}^n [x_i - a_i t, x_i + b_i t]} \left\{ \sum_{i=1}^n V(x_i,t;u_i,a_i,b_i) + \Phi(\boldsymbol{u}) \right\}, \tag{13}$$

for all $\boldsymbol{x}\in\mathbb{R}^n, t\ge 0$, where each function $(x_i,t)\mapsto V(x_i,t;u_i,a_i,b_i)$ on the right-hand side is a continuous function defined explicitly in Chen et al. (2021b).

Let $\boldsymbol{u}^* = (u_1^*,\dots,u_n^*)\in\mathbb{R}^n$ be a minimizer in (13). Note that the minimizer $\boldsymbol{u}^*$ exists since the objective function in the minimization problem in (13) is a lower semi-continuous function with compact domain $\prod_{i=1}^n[x_i - a_i t, x_i + b_i t]$ (see (Rockafellar and Wets, 1998, Theorem 1.9)). However, the minimizer may be not unique. When there are multiple minimizers, let $\boldsymbol{u}^*$ be one such minimizer. Define the trajectory $[0,t]\ni s\mapsto \boldsymbol{\gamma}(s;\boldsymbol{x},t)$ using $\boldsymbol{u}^*$ as

$$\boldsymbol{\gamma}(s;\boldsymbol{x},t) := (\gamma(s;x_1,t,u_1^*,a_1,b_1),\dots,\gamma(s;x_n,t,u_n^*,a_n,b_n)), \tag{14}$$

where the i-th component $\gamma(s;x_i,t,u_i^*,a_i,b_i)$ on the right-hand side is defined explicitly in Chen et al. (2021b). Note that the components of $\boldsymbol{\gamma}(s;\boldsymbol{x},t)$ are independent from each other, and hence, they can be computed in parallel as long as $\boldsymbol{u}^*$ is known.

It is shown in (Chen et al., 2021b, Lemma 3) that the function $u_i\mapsto V(x_i,t;u_i,a_i,b_i)$ is strictly convex and twice continuously differentiable in its domain. Hence, if the initial cost Φ is convex, then (13) is a convex optimization problem that can be numerically solved using convex optimization algorithms. Furthermore, if Φ is quadratic, i.e., Φ is of the form

$$\Phi(\boldsymbol{x}) = \frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \alpha \quad \forall \boldsymbol{x}\in\mathbb{R}^n, \tag{15}$$

for some parameters $\boldsymbol{y}\in\mathbb{R}^n$, $\lambda>0$, and $\alpha\in\mathbb{R}$, then solving the corresponding optimization problem (13) is equivalent to computing the proximal point of the function $\boldsymbol{u}\mapsto \frac{1}{\lambda}\sum_{i=1}^n V(x_i,t;u_i,a_i,b_i)$; this problem can be split into n one-dimensional subproblems, which, in turn, can be solved in parallel and for which explicit formulas were provided in Chen et al. (2021b). In Chen et al. (2021b), we see that when $\lambda=1, \boldsymbol{y}=\boldsymbol{1}$, and $\alpha=0$, this proximal point can be computed on an 11th Gen Intel Laptop Core i7-1165G7 with a 2.80GHz processor in less than 8×10^{-7} seconds on average at one point in 16 dimensions.

Next, we present an efficient algorithm for solving (13) with general convex initial costs Φ. As noted above, in this case, (13) is a convex optimization problem that can be solved using convex optimization algorithms, and the proximal point of the function $\boldsymbol{u}\mapsto\frac{1}{\lambda}\sum_{i=1}^n V(x_i,t;u_i,a_i,b_i)$ can be computed explicitly and in complexity $\Theta(n)$. Thus, proximal point based methods provide a reasonable approach for solving (13). As a demonstration, we show how ADMM (see Glowinski (2014); Boyd et al. (2011)) can be applied to solve (13) in this case. Here, ADMM consists of iterating the following steps:

$$\boldsymbol{v}^{k+1} = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\Phi(\boldsymbol{v}) + \frac{\lambda}{2}\left\|\boldsymbol{v}-\boldsymbol{d}^k+\boldsymbol{w}^k\right\|^2\right\},$$

which is the proximal point of the initial condition,

$$d_i^{k+1} = \operatorname*{arg\,min}_{d_i\in\mathbb{R}}\left\{V(x_i,t;d_i,a_i,b_i) + \frac{\lambda}{2}(v_i^{k+1}-d_i+w_i^k)^2\right\},$$

which is the proximal point of the function $u_i\mapsto\frac{1}{\lambda}V(x_i,t;u_i,a_i,b_i)$, and (6), which is a linear update of coefficients. Once the stopping criteria (7) is met, we set the numerical minimizer to be $\boldsymbol{u}^N = \boldsymbol{d}^N$. We then output the optimal trajectory by:

$$\hat{\boldsymbol{\gamma}}(s;\boldsymbol{x},t) = (\gamma(s;x_1,t,u_1^N,a_1,b_1),\dots,\gamma(s;x_n,t,u_n^N,a_n,b_n)),$$

where the i-th component $\gamma(s;x_i,t,u_i^N,a_i,b_i)$ is defined explicitly in Chen et al. (2021b), and the solution to the HJ PDE by:

$$\hat{V}(\boldsymbol{x},t) = \sum_{i=1}^n V\left(x_i,t;u_i^N,a_i,b_i\right) + \Phi\left(\boldsymbol{u}^N\right),$$

where the i-th component $V\left(x_i,t;u_i^N,a_i,b_i\right)$ in the summation is defined explicitly in Chen et al. (2021b). We note that if $\boldsymbol{v}^{k+1}$ can be computed in complexity $\Theta(n)$, then the overall complexity for each ADMM iteration is also $\Theta(n)$. In other words, using a fixed number of iterations, the curse of dimensionality is avoided in this case. In Chen et al. (2021b), we present several numerical examples that demonstrate the efficiency of our proposed algorithm. As before, all numerical examples were run on an 11th Gen Intel Laptop Core i7-1165G7 with a 2.80GHz processor. For instance, when the initial condition is given by (10), the average time per call of our algorithm to evaluate the solution of the HJ PDE (12) is summarized below for various dimensions n.

n	4	8	12	16
running time (s)	4.6206e-06	3.2871e-05	1.8791e-04	1.9571e-04

Specifically, we see that it takes less than 2×10^{-4} seconds on average to compute the solution at one point in 16 dimensions, which demonstrates that our proposed algorithm can overcome the CoD in this example.

REFERENCES

Akian, M., Bapat, R., and Gaubert, S. (2006). Max-plus algebra. *Handbook of linear algebra*, 39.

Akian, M., Gaubert, S., and Lakhoua, A. (2008). The max-plus finite element method for solving deterministic optimal control problems: basic properties and convergence analysis. *SIAM Journal on Control and Optimization*, 47(2), 817–848.

Bellman, R.E. (1961). *Adaptive control processes: a guided tour*. Princeton university press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1), 1–122.

Chen, P., Darbon, J., and Meng, T. (2021a). Hopf-type representation formulas and efficient algorithms for certain high-dimensional optimal control problems. *arXiv preprint arXiv:2110.02541*.

Chen, P., Darbon, J., and Meng, T. (2021b). Lax-Oleinik-type formulas and efficient algorithms for certain high-dimensional optimal control problems. *arXiv preprint arXiv:2109.14849*.

Coupechoux, M., Darbon, J., Kélif, J., and Sigelle, M. (2019). Optimal trajectories of a uav base station using lagrangian mechanics. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 626–631.

Darbon, J. (2015). On convex finite-dimensional variational methods in imaging sciences and Hamilton–Jacobi equations. *SIAM Journal on Imaging Sciences*, 8(4), 2268–2293.

Darbon, J. and Osher, S. (2016). Algorithms for overcoming the curse of dimensionality for certain Hamilton-Jacobi equations arising in control theory and elsewhere. *Res Math Sci Research in the Mathematical Sciences*, 3(19), 1–26.

Darbon, J. and Meng, T. (2020). On decomposition models in imaging sciences and multi-time Hamilton–Jacobi partial differential equations. *SIAM Journal on Imaging Sciences*, 13(2), 971–1014.

Dower, P.M., McEneaney, W.M., and Zhang, H. (2015). Max-plus fundamental solution semigroups for optimal control problems. In *2015 Proceedings of the Conference on Control and its Applications*, 368–375. SIAM.

Fleming, W. and McEneaney, W. (2000). A max-plus-based algorithm for a Hamilton–Jacobi–Bellman equation of nonlinear filtering. *SIAM Journal on Control and Optimization*, 38(3), 683–710.

Gaubert, S., McEneaney, W., and Qu, Z. (2011). Curse of dimensionality reduction in max-plus based approximation methods: Theoretical estimates and improved pruning algorithms. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, 1054–1061. IEEE.

Glowinski, R. (2014). *On Alternating Direction Methods of Multipliers: A Historical Perspective*, 59–82. Springer Netherlands, Dordrecht.

Li, W. and Todorov, E. (2004). Iterative linear quadratic regulator design for nonlinear biological movement systems. In *ICINCO (1)*, 222–229.

McEneaney, W. (2007). A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs. *SIAM Journal on Control and Optimization*, 46(4), 1239–1276.

McEneaney, W.M. (2006). *Max-plus methods for nonlinear control and estimation*. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA.

McEneaney, W.M., Deshpande, A., and Gaubert, S. (2008). Curse-of-complexity attenuation in the curse-of-dimensionality-free method for HJB PDEs. In *2008 American Control Conference*, 4684–4690. IEEE.

McEneaney, W.M. and Kluberg, L.J. (2009). Convergence rate for a curse-of-dimensionality-free method for a class of HJB PDEs. *SIAM Journal on Control and Optimization*, 48(5), 3052–3079.

Rockafellar, R.T. and Wets, R.J.B. (1998). *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Sideris, A. and Bobrow, J.E. (2005). An efficient sequential linear quadratic algorithm for solving nonlinear optimal control problems. In *Proceedings of the 2005, American Control Conference, 2005.*, 2275–2280 vol. 4.

Yegorov, I. and Dower, P.M. (2017). Perspectives on characteristics based curse-of-dimensionality-free numerical approaches for solving Hamilton–Jacobi equations. *Applied Mathematics & Optimization*, 1–49.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Flag code constructions via group actions ⋆

Xaro Soler-Escrivà *

(Joint work with Miguel Ángel Navarro-Pérez)

* *Department of Mathematics, University of Alacant (Spain) (e-mail: xaro.soler@ua.es).*

Abstract: The aim of this note is to present two specific constructions of flag codes having maximum distance through the action of Singer groups. To this end, we will make use of their transitive action on lines and hyperplanes.

Keywords: Network coding, orbit codes, flag codes, Singer group actions.

1. INTRODUCTION

Random Network Coding is a method for maximize the information rate of an incoherent network which is a directed graph with possibly several senders and receivers. In Koetter and Kschischang (2008) an algebraic framework for such networks is developed by encoding the information in subspaces of a given ambient space $\mathbb{F}_q^n$ over a finite field $\mathbb{F}_q$. In this setting, the set of all subspaces of $\mathbb{F}_q^n$ is turned into a metric space in which two subspaces are closer the greater their intersection. A *constant dimension code* is a set of subspaces of $\mathbb{F}_q^n$ having the same dimension. Since their definition, research works on these codes have proliferated considerably (see for instance Horlemann-Trautmann and Rosenthal (2018) and the references therein). *Flag codes* can be seen as a generalization of constant dimension codes. In this case, the codewords are flags on $\mathbb{F}_q^n$, that is, tuples of nested vector subspaces of prescribed dimensions (Liebhold et al. (2018)). The aim of this note is to explain how we can construct flag codes having maximum distance through the action of Singer groups. Its content is part of the article Navarro-Pérez and Soler-Escrivà (2022), where the reader can find the required proofs. The structure of the note is the following: In the second section we give some preliminaries. In Section 3 we explain how to construct flag codes with maximum distance through the action of a Singer group.

2. PRELIMINARIES

The set of all subspaces of $\mathbb{F}_q^n$ of dimension k is called the *Grassmann variety*, denoted by $\mathcal{G}_q(k,n)$. Given $\mathcal{U},\mathcal{V}\in\mathcal{G}_q(k,n)$, their *subspace distance* is defined as $d_S(\mathcal{U},\mathcal{V}) = \dim(\mathcal{U}+\mathcal{V}) - \dim(\mathcal{U}\cap\mathcal{V}) = 2(k-\dim(\mathcal{U}\cap\mathcal{V}))$. Then, a *constant dimension code* $\mathcal{C}$ is a nonempty subset of $\mathcal{G}_q(k,n)$ and its minimum distance, $d_S(\mathcal{C})$, is the minimum of the distances between distinct pairs of elements of $\mathcal{C}$. The code $\mathcal{C}$ is said to be a *spread code* if it is a spread in the geometrical sense, that is, its elements pairwise intersect trivially and cover the whole space $\mathbb{F}_q^n$. Spreads of dimension k exist if, and only if, k divides n and, in this case, they have cardinality $(q^n-1)/(q^k-1)$.

We put $\mathbb{F}_q^{k\times n}$ for the set of all $k\times n$ matrices with entries in $\mathbb{F}_q$ and $\mathrm{GL}_n(\mathbb{F}_q)$ for the *general linear group* of degree n over $\mathbb{F}_q$, composed by all invertible matrices in $\mathbb{F}_q^{n\times n}$. Given $\mathcal{U}\in\mathcal{G}_q(k,n)$ and a full-rank matrix $U\in\mathbb{F}_q^{k\times n}$ such that $\mathcal{U}=\mathrm{rowsp}(U)$, the operation $\mathcal{U}\cdot A=\mathrm{rowsp}(UA)$ defines a group action of $\mathrm{GL}_n(\mathbb{F}_q)$ on the Grassmann variety, since it is independent from the choice of U (Trautmann et al. (2010)). This way, for a subgroup $\mathbf{H}$ of $\mathrm{GL}_n(\mathbb{F}_q)$, the *orbit code* $\mathrm{O}_{\mathbf{H}}(\mathcal{U})$ is just the constant dimension code arising as the orbit of $\mathcal{U}$ under the action of $\mathbf{H}$. That is, $\mathrm{O}_{\mathbf{H}}(\mathcal{U})=\{\mathcal{U}\cdot A \mid A\in\mathbf{H}\}\subseteq\mathcal{G}_q(k,n)$.

Given integers $0<t_1<\cdots<t_r<n$, a flag $\mathcal{F}=(\mathcal{F}_1,\ldots,\mathcal{F}_r)$ of type $(t_1,\ldots,t_r)$ on $\mathbb{F}_q^n$ is a sequence of nested subspaces of $\mathbb{F}_q^n$, $\mathcal{F}_1\subseteq\cdots\subseteq\mathcal{F}_r$, with $\mathcal{F}_i\in\mathcal{G}_q(t_i,n)$, for all $i=1,\ldots,r$. We say that $\mathcal{F}_i$ is the *i-th subspace* of the flag $\mathcal{F}$ and when the type vector is $(1,2,\ldots,n-1)$ we speak about *full flags*. The subspace distance defined for the Grassmann variety can be naturally extended to flags as follows. Given $\mathcal{F}=(\mathcal{F}_1,\ldots,\mathcal{F}_r)$ and $\mathcal{F}'=(\mathcal{F}'_1,\ldots,\mathcal{F}'_r)$ two flags of type $(t_1,\ldots,t_r)$, their *flag distance* is $d_f(\mathcal{F},\mathcal{F}')=\sum_{i=1}^r d_S(\mathcal{F}_i,\mathcal{F}'_i)$. A *flag code of type* $(t_1,\ldots,t_r)$ is just a nonempty set $\mathcal{C}$ consisting of flags of this type vector and its *minimum distance* is $d_f(\mathcal{C})=\min\{d_f(\mathcal{F},\mathcal{F}')\mid\mathcal{F},\mathcal{F}'\in\mathcal{C},\ \mathcal{F}\neq\mathcal{F}'\}$. For any $i\in\{1,\ldots,r\}$, the *i-projected code* $\mathcal{C}_i$ of $\mathcal{C}$ is defined in Alonso-González et al. (2020) as the constant dimension code $\mathcal{C}_i=\{\mathcal{F}_i\mid(\mathcal{F}_1,\ldots,\mathcal{F}_r)\in\mathcal{C}\}\subseteq\mathcal{G}_q(t_i,n)$. Notice that $d_f(\mathcal{C})\leq 2\left(\sum_{2t_i\leq n}t_i+\sum_{2t_i>n}(n-t_i)\right)$. We say that $\mathcal{C}$ is an *optimum distance flag code* (ODFC) whenever $d_f(\mathcal{C})$ attains this upper bound. It follows that the projected codes of an ODFC are constant dimension codes of maximum distance. However, this is not a sufficient condition in order to obtain ODFC (Alonso-González et al. (2020)).

Clearly, the action of the general linear group on the Grassmann variety can be extended to flags. Given a flag $\mathcal{F}=(\mathcal{F}_1,\ldots,\mathcal{F}_r)$ and a subgroup $\mathbf{H}$ of $\mathrm{GL}(n,q)$, the *orbit flag code* generated by $\mathcal{F}$ under the action of $\mathbf{H}$ is

$$\mathrm{O}_{\mathbf{H}}(\mathcal{F})=\{\mathcal{F}\cdot A\mid A\in\mathbf{H}\}=\{(\mathcal{F}_1\cdot A,\ldots,\mathcal{F}_r\cdot A)\mid A\in\mathbf{H}\}.$$

⋆ The authors received financial support of Ministerio de Ciencia e Innovación (PID2019-108668GB-I00). The first author is supported by Generalitat Valenciana and Fondo Social Europeo (Grant number: ACIF/2018/196).

Putting $\mathrm{Stab}_{\mathbf{H}}(\mathcal{F}) = \{A \in \mathbf{H} \mid \mathcal{F}\cdot A = \mathcal{F}\}$ for the stabilizer subgroup of this action, it holds $|\mathrm{O}_{\mathbf{H}}(\mathcal{F})| = \frac{|\mathbf{H}|}{|\mathrm{Stab}_{\mathbf{H}}(\mathcal{F})|}$. Moreover, the minimum distance of the orbit flag code is

$$d_f(\mathrm{O}_{\mathbf{H}}(\mathcal{F})) = \min\{d_f(\mathcal{F}, \mathcal{F}\cdot A) \mid A \in \mathbf{H} \setminus \mathrm{Stab}_{\mathbf{H}}(\mathcal{F})\}$$

and it holds $d_f(\mathrm{O}_{\mathbf{H}}(\mathcal{F})) = 0$ if, and only if, $\mathrm{Stab}_{\mathbf{H}}(\mathcal{F}) = \mathbf{H}$. The projected codes of an orbit flag code are orbit (subspace) codes. More precisely, for every $1 \leq i \leq r$, we have $\mathrm{O}_{\mathbf{H}}(\mathcal{F})_i = \mathrm{O}_{\mathbf{H}}(\mathcal{F}_i) \subseteq \mathcal{G}_q(t_i, n)$. Besides, the stabilizer subgroup of $\mathcal{F}$ is closely related to the ones of its subspaces: $\mathrm{Stab}_{\mathbf{H}}(\mathcal{F}) = \bigcap_{i=1}^{r} \mathrm{Stab}_{\mathbf{H}}(\mathcal{F}_i)$ (Alonso-González et al. (2021b)).

Remark that, fixed an acting subgroup $\mathbf{H}$ of $\mathrm{GL}_n(\mathbb{F}_q)$, the cardinality of the flag code $\mathrm{O}_{\mathbf{H}}(\mathcal{F})$ and their projected codes are determined by the orders of the corresponding stabilizer subgroups. Since $\mathrm{Stab}_{\mathbf{H}}(\mathcal{F}) = \mathrm{Stab}_{\mathrm{GL}_n(\mathbb{F}_q)}(\mathcal{F}) \cap \mathbf{H}$, one way to construct orbit flag codes is to start from a subspace whose stabilizer in $\mathrm{GL}_n(\mathbb{F}_q)$ is known and consider subgroups $\mathbf{H}$ that trivially intersect it. This is the approach of the constructions given in Liebhold et al. (2018). More generally, we can look for subgroups of $\mathrm{GL}_n(\mathbb{F}_q)$ whose action on certain subspaces of $\mathbb{F}_q^n$ is known and try to extend this action to flags in an appropriate manner. This is our strategy, using classical results about the action of Singer groups on lines and hyperplanes (Theorem 1).

3. ORBIT FLAG CODES FROM SINGER GROUPS

The aim of this section is to present two constructions of ODFC. In Subsection 3.1 we construct ODFCs on $\mathbb{F}_q^n$ having a k-spread as a projected code, for k a divisor of n. To do so, we consider flags of type $(1, \ldots, k, n-k, \ldots, n-1)$. Such a construction leads to full flag codes whenever $n = 2k$ or $k = 1$ and $n = 3$. In Subsection 3.2, we build ODFCs of full type vector for the remaining cases. Both constructions are closely related to Singer groups. A *Singer subgroup* of $\mathrm{GL}_n(\mathbb{F}_q)$ is a cyclic subgroup of the largest possible order, which is $q^n - 1$. Singer subgroups form a conjugacy class of $\mathrm{GL}_n(\mathbb{F}_q)$ and act on the Grassmann varieties of lines and hyperplanes as follow:

Theorem 1. (Beth et al., 1999, Th. 6.2) Any Singer cyclic subgroup $\mathbf{S}$ of $\mathrm{GL}_n(\mathbb{F}_q)$ acts transitively on both $\mathcal{G}_q(1, n)$ and $\mathcal{G}_q(n-1, n)$. Moreover, for any $l \in \mathcal{G}_q(1, n)$ and any $h \in \mathcal{G}_q(n-1, n)$, it holds that $\mathrm{Stab}_{\mathbf{S}}(l) = \mathrm{Stab}_{\mathbf{S}}(h) = \{aI_n \mid a \in \mathbb{F}_q^*\}$, which is the unique cyclic subgroup of $\mathbf{S}$ of order $q-1$.

3.1 Orbit ODFC from spreads

Assume that $n = ks$, for $s \geq 2$, and that $\mathcal{C}$ is an ODFC such that dimension k appears in its type vector. Then $|\mathcal{C}| \leq \frac{q^n-1}{q^k-1}$ and the equality holds if, and only if, the projected code $\mathcal{C}_i$ is a k-spread of $\mathbb{F}_q^n$ (Alonso-González et al. (2020)). Moreover, $(1, \ldots, k, n-k, \ldots, n-1)$ is the largest type vector for an ODFC on $\mathbb{F}_q^n$ having a k-spread as one of its projected codes (Alonso-González et al. (2021a)).

Motivated by these results, we consider for our construction an arbitrary flag of the *full admissible type vector* $(1, \ldots, k, n-k, \ldots, n-1)$. For sake of simplicity, we will denote it by $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_k, \mathcal{F}_{n-k}, \ldots, \mathcal{F}_{n-1})$. Then the following result holds:

Theorem 2. (Navarro-Pérez and Soler-Escrivà, 2022, Th. 4.11) Given a subgroup $\mathbf{H}$ of $\mathrm{GL}_n(\mathbb{F}_q)$ and a flag $\mathcal{F}$ of the full admissible type vector, the following sentences are equivalent:

(*i*) $\mathrm{O}_{\mathbf{H}}(\mathcal{F})$ is an ODFC.
(*ii*) $\mathrm{Stab}_{\mathbf{H}}(\mathcal{F}_k) = \mathrm{Stab}_{\mathbf{H}}(\mathcal{F}_{n-k}) \subseteq \mathrm{Stab}_{\mathbf{H}}(\mathcal{F}_i)$, for every i and the subspace codes $\mathrm{O}_{\mathbf{H}}(\mathcal{F}_k)$ and $\mathrm{O}_{\mathbf{H}}(\mathcal{F}_{n-k})$ are of maximum distance.

Now, the idea is to use Theorem 1 in order to achieve the condition (*ii*) of Theorem 2. To do so, let ω be a primitive element of the field $\mathbb{F}_{q^k}$ and $M_k \in \mathrm{GL}_k(\mathbb{F}_q)$ the companion matrix associated to the primitive polynomial of ω over $\mathbb{F}_q$. It turns out that $\mathbb{F}_{q^k} \cong \mathbb{F}_q[\omega] \cong \mathbb{F}_q[M_k]$, where the last field isomorphism is given by $\phi(\sum_{i=0}^{k-1} a_i\omega^i) = \sum_{i=0}^{k-1} a_i M_k^i$. In particular $\phi(\omega) = M_k$ and the multiplicative order of M_k is $q^k - 1$. Thus, $\langle M_k \rangle$ is a Singer subgroup of $\mathrm{GL}_k(\mathbb{F}_q)$. Equivalently, M_k is a primitive element of the finite field $\mathbb{F}_q[M_k] \subseteq \mathbb{F}_q^{k\times k}$.

The field isomorphism ϕ is useful to map vector subspaces of $\mathbb{F}_{q^k}^s$ into vector subspaces of $\mathbb{F}_q^n$. For $m \in \{1, \ldots, s\}$, one has the embedding $\varphi : \mathcal{G}_{q^k}(m, s) \longrightarrow \mathcal{G}_q(km, n)$ given by

$$\mathrm{rowsp}\begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{ms} \end{pmatrix} \longmapsto \mathrm{rowsp}\left(\begin{array}{c|c|c} \phi(a_{11}) & \cdots & \phi(a_{1s}) \\ \hline \vdots & \ddots & \vdots \\ \hline \phi(a_{m1}) & \cdots & \phi(a_{ms}) \end{array}\right),$$

which is called a *field reduction map*. In particular, φ preserves intersections of subspaces, since it is injective. Therefore, an m-spread of $\mathbb{F}_{q^k}^s$ will be mapped into a km-spread of $\mathbb{F}_q^n$. We will use two constant dimension codes of $\mathbb{F}_q^n$ constructed in this way. First, from the spread of lines of $\mathbb{F}_{q^k}^s$, we consider

$$\mathcal{S} = \varphi(\mathcal{G}_{q^k}(1, s)) \subseteq \mathcal{G}_q(k, n),$$

which is a k-spread of $\mathbb{F}_q^n$. Originally, due to Segre, in the network coding setting, this construction appears for the first time in Manganiello et al. (2008). Secondly, from the Grassmannian of hyperplanes of $\mathbb{F}_{q^k}^s$, we obtain

$$\mathcal{H} = \varphi(\mathcal{G}_{q^k}(s-1, s)) \subseteq \mathcal{G}_q(n-k, n),$$

which is a constant dimension code of $\mathbb{F}_q^n$ with maximum distance.

On the other hand, we can also use ϕ to obtain the following group monomorphism:

$$\begin{array}{rccc} \psi: & \mathrm{GL}_s(\mathbb{F}_{q^k}) & \longrightarrow & \mathrm{GL}_{ks}(\mathbb{F}_q) \\ & \begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \cdots & a_{ss} \end{pmatrix} & \longmapsto & \left(\begin{array}{c|c|c} \phi(a_{11}) & \cdots & \phi(a_{1s}) \\ \hline \vdots & \ddots & \vdots \\ \hline \phi(a_{s1}) & \cdots & \phi(a_{ss}) \end{array}\right). \end{array}$$

In this way, we can establish the following relation between the group action of $\mathrm{GL}_s(\mathbb{F}_{q^k})$ on $\mathcal{G}_{q^k}(m, s)$ and the group action of $\mathrm{GL}_n(\mathbb{F}_q)$ on $\mathcal{G}_q(km, n)$:

$$\varphi(\mathcal{V}\cdot A) = \varphi(\mathcal{V})\cdot\psi(A), \tag{1}$$

for all $\mathcal{V} \in \mathcal{G}_{q^k}(m, s)$ and $A \in \mathrm{GL}_s(\mathbb{F}_{q^k})$. In particular, we will use this equality to relate the respective actions of two Singer subgroups in which we are very interested.

Let α be a primitive element of $\mathbb{F}_{q^n}$ and $M_s \in \mathrm{GL}_s(\mathbb{F}q^k)$ the companion matrix of the minimal polynomial of α

over $\mathbb{F}_{q^k}$, then $\mathbb{F}_{q^n} \cong \mathbb{F}_{q^k}[\alpha] \cong \mathbb{F}_{q^k}[M_s]$. Therefore, the multiplicative order of M_s is q^n-1 and $\langle M_s\rangle$ is a Singer subgroup of $\mathrm{GL}_s(\mathbb{F}_{q^k})$. Besides, it turns out that $\psi(\langle M_s\rangle) = \langle \psi(M_s)\rangle$ is a Singer subgroup of $\mathrm{GL}_n(\mathbb{F}_q)$. Given a line $l_0 \in \mathcal{G}_{q^k}(1,s)$, put $\varphi(l_0) = \mathcal{S}_0 \in \mathcal{S}$. In accordance with Theorem 1 and (1) we can write

$$\mathcal{S} = \varphi(\mathcal{G}_{q^k}(1,s)) = \varphi(\mathrm{O}_{\langle M_s\rangle}(l_0)) = \{\varphi(l_0 \cdot A) \mid A \in \langle M_s\rangle\}$$
$$= \{\mathcal{S}_0 \cdot \psi(A) \mid \psi(A) \in \langle \psi(M_s)\rangle\} = \mathrm{O}_{\langle \psi(M_s)\rangle}(\mathcal{S}_0).$$

In an analogous way, given a hyperplane $h_0 \in \mathcal{G}_{q^k}(s-1,s)$, denote $\varphi(h_0) = \mathcal{H}_o \in \mathcal{H}$. Then, we obtain that

$$\mathcal{H} = \varphi(\mathcal{G}_{q^k}(s-1,s)) = \varphi(\mathrm{O}_{\langle M_s\rangle}(h_0)) = \mathrm{O}_{\langle \psi(M_s)\rangle}(\mathcal{H}_0).$$

That is, the transitive action of $\langle M_s\rangle$ on the lines and hyperplanes of $\mathbb{F}_{q^k}^s$ is translated into the transitive action of $\langle \psi(M_s)\rangle$ on the constant dimension codes of maximum distance $\mathcal{S}$ and $\mathcal{H}$. Moreover, for any $\mathcal{S}_0 \in \mathcal{S}$ and $\mathcal{H}_0 \in \mathcal{H}$, Theorem 1 also leads to

$$\mathrm{Stab}_{\langle \psi(M_s)\rangle}(\mathcal{S}_0) = \mathrm{Stab}_{\langle \psi(M_s)\rangle}(\mathcal{H}_0) = \{\psi(aI_s) \mid a \in \mathbb{F}_{q^k}^*\},$$

which has order q^k-1.

With all these ingredients, we can now apply Theorem 2 in order to characterize those subgroups of $\langle \psi(M_s)\rangle$ that are appropriate to construct ODFCs.

Theorem 3. Let $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_k, \mathcal{F}_{n-k}, \ldots, \mathcal{F}_{n-1})$ be a flag of full admissible type vector such that $\mathcal{F}_k \in \mathcal{S}$ and $\mathcal{F}_{n-k} \in \mathcal{H}$. For any positive integer t dividing q^n-1, consider the unique subgroup $\mathbf{T}$ of $\langle \psi(M_s)\rangle$ of order t. Then:

(*i*) $|\mathrm{O}_{\mathbf{T}}(\mathcal{F})| = \frac{t}{\gcd(t,q-1)}$.
(*ii*) $\mathrm{O}_{\mathbf{T}}(\mathcal{F})$ is an ODFC if, and only if, $\gcd(t,q^k-1) = \gcd(t,q-1) \neq t$.

Theorem 3 states which subgroups of $\langle \psi(M_s)\rangle$ allow the construction of ODFCs as a single orbit of them. Notice that bigger subgroups not always will provide bigger orbit flag codes. In addition, it may eventually happen that some subgroup provides an orbit ODFC of the maximum possible size, that is, $\frac{q^n-1}{q^k-1}$. Otherwise, we will consider unions of orbits under the action of that subgroup in order to obtain an optimal construction (Navarro-Pérez and Soler-Escrivà, 2022, Th. 4.13). Clearly, the larger the size of each orbit, the fewer orbits we need to reach the maximum size and vice versa. In particular, the degenerate case where an ODFC is constructed as a union of $\frac{q^n-1}{q^k-1}$ orbits with just one element is also contemplated. All these considerations are reflected in the following examples, in which we apply Theorem 3 for different values of the parameters.

Examples. With the notation of Theorem 3, we consider all the divisors t of q^n-1 such that $\gcd(t,q^k-1) = \gcd(t,q-1)$ and the corresponding subgroup $\mathbf{T}$ of $\langle \psi(M_s)\rangle$ of order t. Consider a flag $\mathcal{F}$ of the full admissible type $(1,\ldots,k,n-k,\ldots,n-1)$ such that $\mathcal{F}_k \in \mathcal{S}$ and $\mathcal{F}_{n-k} \in \mathcal{H}$. Finally, denote by m the number of required orbits of $\mathbf{T}$ to attain the maximum size, $\frac{q^n-1}{q^k-1}$, for an ODFC with these parameters.

(1) Put $q=3$, $k=3$ and $n=6$. Thus, $k=n-k$, $q^n-1=728$, $q^k-1=26$ and $\frac{q^n-1}{q^k-1}=28$. Then

t	1	2	4	7	8	14	28	56
$\lvert\mathrm{O}_{\mathbf{T}}(\mathcal{F})\rvert$	1	1	2	7	4	7	14	28
m	28	28	14	4	7	4	2	1

Notice that, in this case, the subgroup of order $t=56$ allows us to obtain ODFCs of full type vector and having the best possible size, i.e., 28, by using a single orbit. Moreover, remark that the subgroup of order $t=8$ gives an orbit ODFC of smaller size than the obtained with the subgroup of order $t=7$.

(2) Put $q=4$, $k=3$ and $n=9$. Thus, $n-k=6$, $q^n-1=262143$, $q^k-1=63$ and $\frac{q^n-1}{q^k-1}=4161$. Then

t	1	3	19	57	73	219	1387	4161
$\lvert\mathrm{O}_{\mathbf{T}}(\mathcal{F})\rvert$	1	1	19	19	73	73	1387	1387
m	4161	4161	219	219	57	57	3	3

The largest orbit size is 1387 and it is obtained when the acting group has order either 1387 or 4161. On the other hand, the maximum possible size of an ODFC with these parameters is 4161. Hence, in order to achieve that cardinality, we must consider the union of, at least, 3 different orbits.

The orbital constructions of ODFC provided in this section present a restriction on the type vector, coming from the condition of having a spread as a projected code. However, there are two possible situations in which flag codes of full type can be given by using Theorem 3. First, for even values of n, taking the divisor $k=\frac{n}{2}$ leads to a construction of full type in which $k=n-k$. This particular case was first studied in Alonso-González et al. (2021b), where a construction using the action of a Singer subgroup of $\mathrm{SL}_{2k}(\mathbb{F}_q)$ is presented. On the other hand, for $n=3$ and $k=1$, the action of a Singer subgroup of $\mathrm{GL}_3(\mathbb{F}_q)$ on the Grassmannian of lines and hyperplanes gives us a construction of type $(1,2)$. This construction is also known and appears in (Kurz, 2021, Prop. 2.5), where the author shows that it is the one with the biggest cardinality among ODFC of full type when $n=3$. In the following section, we consider the remaining situations, that is, we address the construction of orbit ODFCs of full type vector on $\mathbb{F}_q^n$ for odd values of $n>3$.

3.2 Orbit ODFC of full type vector

In this section, we assume that $n=2k+1$, for some $k>1$. In this case, instead of Theorem 2, the following result is achieved.

Theorem 4. (Navarro-Pérez and Soler-Escrivà, 2022, Th. 4.11) Given a subgroup $\mathbf{H}$ of $\mathrm{GL}_n(\mathbb{F}_q)$ and a full flag $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_{2k})$ on $\mathbb{F}_q^n$, the following sentences are equivalent:

(*i*) $\mathrm{O}_{\mathbf{H}}(\mathcal{F})$ is an ODFC.
(*ii*) $\mathrm{Stab}_{\mathbf{H}}(\mathcal{F}_k) = \mathrm{Stab}_{\mathbf{H}}(\mathcal{F}_{k+1}) \subseteq \mathrm{Stab}_{\mathbf{H}}(\mathcal{F}_i)$, for every i and the subspace codes $\mathrm{O}_{\mathbf{H}}(\mathcal{F}_k)$ and $\mathrm{O}_{\mathbf{H}}(\mathcal{F}_{k+1})$ are of maximum distance.

Our aim is to use the previous theorem in order to obtain a specific construction of orbit ODFC of full type vector. For this, let us consider $M_{k+1} \in \mathrm{GL}_{k+1}(\mathbb{F}_q)$ the companion matrix of a primitive polynomial of degree $k+1$ in $\mathbb{F}_q[x]$. Then $\langle M_{k+1}\rangle$ is a Singer subgroup of $\mathrm{GL}_{k+1}(\mathbb{F}_q)$ and

$\mathbb{F}_q[M_{k+1}]$ is a matrix representation of the finite field of q^{k+1} elements. Let us write

$$g = \left(\begin{array}{c|c} I_k & 0_{k\times(k+1)} \\ \hline 0_{(k+1)\times k} & M_{k+1} \end{array}\right) \in \mathrm{GL}_n(\mathbb{F}_q)$$

and consider the cyclic group

$$\mathbf{G} = \langle g \rangle = \left\{ g^i \mid 0 \le i \le q^{k+1} - 2 \right\}.$$

Clearly, $\mathbf{G}$ is a subgroup of order $q^{k+1}-1$ of $\mathrm{GL}_n(\mathbb{F}_q)$, isomorphic to the Singer subgroup $\langle M_{k+1} \rangle$ of $\mathrm{GL}_{k+1}(\mathbb{F}_q)$. In the rest of this section, the orbit codes considered will be always generated by the action of this specific group $\mathbf{G}$.

We start by characterizing the subspaces of dimensions k and $k+1$ of $\mathbb{F}_q^n$ whose orbits under the action of $\mathbf{G}$ are constant dimension codes of maximum distance. Given arbitrary subspaces $\mathcal{U} = \mathrm{rowsp}(U) \in \mathcal{G}_q(k,n)$ and $\mathcal{V} = \mathrm{rowsp}(V) \in \mathcal{G}_q(k+1,n)$, their respective full-rank generator matrices can split into two blocks as

$$U = (U_1 \mid U_2) \quad \text{and} \quad V = (V_1 \mid V_2), \tag{2}$$

where U_1 (resp. V_1) denotes the first k columns of U (resp. V). Therefore, $U_1 \in \mathbb{F}_q^{k\times k}$, $U_2 \in \mathbb{F}_q^{k\times(k+1)}$, $V_1 \in \mathbb{F}_q^{(k+1)\times k}$ and $V_2 \in \mathbb{F}_q^{(k+1)\times(k+1)}$. Now, consider the orbit codes

$$\mathrm{O}_{\mathbf{G}}(\mathcal{U}) = \{\mathcal{U} \cdot g^i \mid 0 \le i \le q^{k+1} - 2\} \tag{3}$$

and

$$\mathrm{O}_{\mathbf{G}}(\mathcal{V}) = \{\mathcal{V} \cdot g^i \mid 0 \le i \le q^{k+1} - 2\}. \tag{4}$$

With the notation of (2) the following results hold.

Proposition 5. The orbit code $\mathrm{O}_{\mathbf{G}}(\mathcal{U})$ defined in (3) has maximum distance if, and only if, $\mathrm{rk}(U_1) = \mathrm{rk}(U_2) = k$. Its cardinality is $|\mathrm{O}_{\mathbf{G}}(\mathcal{U})| = |\mathbf{G}| = q^{k+1} - 1$.

Proposition 6. The orbit code $\mathrm{O}_{\mathbf{G}}(\mathcal{V})$ defined in (4) has maximum distance if, and only if, $\mathrm{rk}(V_1) = k$ and $\mathrm{rk}(V_2) = k+1$. Its size is $|\mathrm{O}_{\mathbf{G}}(\mathcal{V})| = |\mathbf{G}| = q^{k+1} - 1$.

In the following, we use the previous characterizations for constant dimension codes of maximum distance in order to provide orbit ODFCs of full type on $\mathbb{F}_q^n$. To do so, we need to consider nested subspaces $\mathcal{U} \subsetneq \mathcal{V}$ of dimensions k and $k+1$, respectively. Using the notation of (2), we can formulate the problem in a matrix approach: given a full-rank generator matrix $U = (U_1 \mid U_2) \in \mathbb{F}_q^{k\times n}$ of $\mathcal{U}$, we consider a subspace $\mathcal{V}$ spanned by the rows of a matrix $V \in \mathbb{F}_q^{(k+1)\times n}$, obtained by adding an appropriate row to U. In other words, we choose vectors $\mathbf{v}_1 \in \mathbb{F}_q^k$ and $\mathbf{v}_2 \in \mathbb{F}_q^{k+1}$ such that the matrix

$$V = (V_1 \mid V_2) = \left(\begin{array}{c|c} U_1 & U_2 \\ \hline \mathbf{v}_1 & \mathbf{v}_2 \end{array}\right) \tag{5}$$

has rank equal to $k+1$. Using this notation, we present the next construction of ODFCs arising from the action of the group $\mathbf{G}$ defined in this section.

Theorem 7. Let $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_{2k})$ be a full flag on $\mathbb{F}_q^n$ such that $\mathcal{F}_k = \mathcal{U} = \mathrm{rowsp}(U_1 \mid U_2)$ and $\mathcal{F}_{k+1} = \mathcal{V} = \mathrm{rowsp}(V_1 \mid V_2)$, with generator matrix as in (5). The following statements are equivalent:

(i) The flag code $\mathrm{O}_{\mathbf{G}}(\mathcal{F})$ is an ODFC.
(ii) The matrices $U_1 \in \mathbb{F}_q^{k\times k}$ and $V_2 \in \mathbb{F}_q^{(k+1)\times(k+1)}$ are invertible.

In this situation, $|\mathrm{O}_{\mathbf{G}}(\mathcal{F})| = |\mathbf{G}| = q^{k+1} - 1$.

The ODFC constructed in Theorem 7 contains $q^{k+1} - 1$ flags. It is the largest size for orbits under the action of the group $\mathbf{G}$. On the other hand, as proved in (Kurz, 2021, Prop. 2.4), the maximum possible cardinality for ODFCs of full type on $\mathbb{F}_q^n$ is exactly $q^{k+1}+1$. Consequently, our orbital construction is only two flags away from reaching the mentioned bound. In order to complete this construction into an ODFC with size $q^{k+1}+1$, we form the following subspaces

$$\begin{aligned} &\mathcal{U}' = \mathrm{rowsp}(U_1 \mid 0_{k\times(k+1)}), \ \mathcal{V}' = \mathrm{rowsp}\left(\begin{array}{c|c} U_1 & 0_{k\times(k+1)} \\ \hline \mathbf{v}_1 & \mathbf{v}_2 \end{array}\right), \\ &\mathcal{U}'' = \mathrm{rowsp}(0_{k\times k} \mid U_2), \qquad \mathcal{V}'' = \mathrm{rowsp}\left(\begin{array}{c|c} 0_{k\times k} & U_2 \\ \hline \mathbf{v}_1 & \mathbf{v}_2 \end{array}\right). \end{aligned} \tag{6}$$

With this notation, as long as $\mathrm{rk}(U_1) = \mathrm{rk}(U_2) = k$ and $\mathbf{v}_2 \notin \mathrm{rowsp}(U_2)$, the next result holds.

Theorem 8. Let $\mathcal{F}, \mathcal{F}', \mathcal{F}''$ be full flags on $\mathbb{F}_q^n$ such that $\mathcal{F}_k = \mathcal{U}$ and $\mathcal{F}_{k+1} = \mathcal{V}$ defined as in *Theorem 7* and

$$\mathcal{F}'_k = \mathcal{U}', \ \mathcal{F}'_{k+1} = \mathcal{V}', \ \mathcal{F}''_k = \mathcal{U}'', \ \mathcal{F}''_{k+1} = \mathcal{V}''.$$

Then the flag code $\mathcal{C} = \mathrm{O}_{\mathbf{G}}(\mathcal{F}) \cup \{\mathcal{F}', \mathcal{F}''\}$ is an ODFC with the maximum possible cardinality, i.e., $q^{k+1}+1$, if, and only if, $\mathbf{v}_1 = \mathbf{0}_k$.

REFERENCES

Alonso-González, C., Navarro-Pérez, M.A., and Soler-Escrivà, X. (2020). Flag codes from planar spreads in network coding. *Finite Fields and Their Applications*, 68, 101745.

Alonso-González, C., Navarro-Pérez, M.A., and Soler-Escrivà, X. (2021a). Optimum distance flag codes from spreads via perfect matchings in graphs. *Journal of Algebraic combinatorics*, 54, 1279–1297.

Alonso-González, C., Navarro-Pérez, M.A., and Soler-Escrivà, X. (2021b). An orbital construction of optimum distance flag codes. *Finite Fields and Their Applications*, 73, 101861.

Beth, T., Jungnickel, D., and Lenz, H. (1999). *Design Theory*, volume 1 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 2 edition.

Horlemann-Trautmann, A.L. and Rosenthal, J. (2018). Constructions of constant dimension codes. In *Network Coding and Subspace Designs*, 25–42. Springer.

Koetter, R. and Kschischang, F.R. (2008). Coding for errors and erasures in random network coding. *IEEE Transactions on Information theory*, 54(8), 3579–3591.

Kurz, S. (2021). Bounds for flag codes. *Designs, Codes and Cryptography*, 89, 2759–2785.

Liebhold, D., Nebe, G., and Vazquez-Castro, A. (2018). Network coding with flags. *Designs, Codes and Cryptography*, 86(2), 269–284.

Manganiello, F., Gorla, E., and Rosenthal, J. (2008). Spread codes and spread decoding in network coding. In *2008 IEEE International Symposium on Information Theory*, 881–885.

Navarro-Pérez, M.Á. and Soler-Escrivà, X. (2022). Flag codes of maximum distance and constructions using singer groups. *Finite Fields and Their Applications*, 80, 102011.

Trautmann, A.L., Manganiello, F., and Rosenthal, J. (2010). Orbit codes, a new concept in the area of network coding. In *2010 IEEE Information Theory Workshop*, 1–4. IEEE.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Towards a reduction of the public-key size of a Gabidulin codes based encryption scheme

Pierre Loidreau *

* *DGA and Institut de Recherche Mathmatique de Rennes, IRMAR - UMR CNRS 6625*

Abstract: We exhibit a way to reduce the public-key size of a rank metric based public-key cryptosystem. This approach does not use a structural property of the code but exhibit some particular generator matrices that have a quasi-cyclic like structure.

Keywords: Post-Quantum Cryptography, Coding Theory, Rank metric based cryptosystems

1. INTRODUCTION

Code-based cryptography is among the main solutions for post-quantum cryptography together with lattice-based and isogeny-based cryptography. A McEliece type OW-PKE based on Gabidulin codes was proposed in Loidreau (2017). It mimics the McEliece type OW-PKE proposed in Hamming metric, but compared to modern PKE and especially those proposed at the NIST post-quantum standardization process this one has three main advantages.

- Decryption is deterministic. This makes much easier to handle as an IND-CCA version than Lattice based schemes and MDPC codes based schemes.
- Key size is between one and two orders of magnitude smaller than other Hamming metric based cryptosystem. It favourably compares to unstructured lattice-based PKEs such as the one used in `FrodoKEM`.
- The cipher text is small compared to unstructured lattice and compares favourably with structured lattices.

On the other hand, and this forms the main drawback, the security analysis is not yet sufficiently stabilized. The security of the scheme relies on two paradigms:

- The complexity of distinguishing the public code from random
- The complexity of decoding a random code in rank metric.

The latter problem was proven to be a hard problem in the complexity class ZPP Gaborit and Zémor (2015). However, in practice significant recent progresses were made in the computational complexity which makes necessary to reconsider the parameters, Bardet et al. (2020). This implies in particuler increasing the public-key size which is the main drawback in using code-based cryptography in limited resources devices.

In our work we propose a new direction towards the reduction of the public-key.

2. BACKGROUND ON GABIDULIN CODES AND ON RANK METRIC

First we define the rank norm of a vector inducing a metric on a finite field. Let q be the power of some prime number.

Definition 1. (Rank of a vector). Let $\mathbb{F}_{q^m}$ be the finite field with q^m elements. Let $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{F}_{q^m}^n$. Then the rank of $\mathbf{a}$ denoted by $\mathrm{Rk}(\mathbf{a})$ is the dimension of the $\mathbb{F}_q$-dimensional vector subspace of $\mathbb{F}_{q^m}$ generated by the components of $\mathbf{a}$, *i.e.*

$$\mathrm{Rk}(\mathbf{a}) \stackrel{def}{=} \dim \langle a_1, \ldots, a_n \rangle_{\mathbb{F}_q}$$

As extremal object in Bose-Mesner algebra Gabidulin codes were first discovered by Delsarte Delsarte (1978). Some years later Gabidulin presented an algebraic theory as well as a polynomial-time decoding algorithm Gabidulin (1985).

Definition 2. Let $0 < k < n \leq m$, and $\mathbf{g} = (g_1, \ldots, g_n) \in \mathbb{F}_{q^m}^n$ of rank n. Then, the k-dimensional Gabidulin code with support vector $\mathbf{g}$ denoted by $\mathcal{G}_k(\mathbf{g})$ is

$$\mathcal{G}_k(\mathbf{g}) = \left\{ \mathbf{x} \left(g_j^{[i]} \right)_{i=0,j=1}^{k-1,n} \mid \mathbf{x} \in \mathbb{F}_{q^m}^k \right\},$$

where $[i] \stackrel{def}{=} q^i$

We consider now more specific Gabidulin codes. Let $u, \lambda, k \in \mathbb{N} \setminus \{0\}$. Let us define $m = \mu u$ and $k = \kappa u$, with $\kappa < \mu$. Let $g_\ell \in \mathbb{F}_{q^m}$, for $\ell = 1, \ldots, u$ such that

$$\mathrm{Rk}(\mathbf{g} = (g_\ell^{[uj]},\ j = 0, \ldots, \mu - 1,\ \ell = 1, \ldots, u)) = m \quad (1)$$

A generator $\mathbf{G}$ matrix of $\mathcal{G}_k(\mathbf{g})$ has the form

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1u} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{u1} & \cdots & \mathbf{G}_{uu} \end{pmatrix} \quad (2)$$

where

$$\forall \ell, s = 1, \ldots, u,\ \mathbf{G}_{s\ell} = \left(g_\ell^{[(s-1)+u(i+j)]} \right)_{i=0,j=0}^{\kappa-1,\mu-1}$$

In other terms the matrices $\mathbf{G}_{s\ell}$ are submatrices formed by the κ first rows of the circulant $\mu \times \mu$ matrices

$$\left(g_\ell^{[(s-1)+u(i+j)]} \right)_{i=0,j=0}^{\mu-1,\mu-1}$$

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Definition 3. A generator matrix of a Gabidulin code $\mathcal{G}_k(\mathbf{g})$ under the form (2) is said to be under quasi-cyclic form.

3. CONSTRUCTION OF GABIDULIN CODES UNDER QUASI-CYCLIC FORM

We show how to find a vector $\mathbf{g}$ or rank m as described in equation (1). Let $g \in \mathbb{F}_{2^m}$ and let us define

$$\mathcal{V}_g \stackrel{def}{=} \langle g, g^{[u]}, \ldots, g^{[u(\mu-1)]} \rangle_{\mathbb{F}_2}$$

Note that we can show that if $h \in \mathbb{F}_{2^m}$ and $h \notin \mathcal{V}_g$ then $\mathcal{V}_g \cap \mathcal{V}_h = \{0\}$

(1) Choose $g_1 \in \mathbb{F}_{2^m}$ such that $\dim(\mathcal{V}_{g_1}) = \mu$.
(2) For $\ell = 2, \ldots, u$ choose $g_\ell \in \mathbb{F}_{2^m} \setminus \cup_{i=1}^{\ell-1} \mathcal{V}_{g_i}$ such that $\dim(\cup_{i=1}^{\ell} \mathcal{V}_{g_i}) = \mu$.

Note that we did not check the effective number of such vectors. At least any normal element of the finite field can give rise to such vectors. Our simulations in MAGMA confirm for $q = 2$ that a very small number of tries is necessary to find a target vector (typically less than 10).

The key proposition is:

Proposition 1. Let $\mathbf{G} = (\mathbf{G}_{s\ell})$ be the generator matrix of the Gabidulin code $\mathcal{G}_k(\mathbf{g})$ under quasi-circulant form where $\mathbf{G}_{s\ell} \in \mathbb{F}_{q^m}^{\kappa \times \mu}$ are $\kappa \times \mu$ submatrices of circulant $\mu \times \mu$ matrices. Let $\mathbf{T} = (\mathbf{T}_{s\ell})$ where $\mathbf{T}_{s\ell} \in \mathbb{F}_{q^m}^{\mu \times \mu}$ are circulant $\mu \times \mu$-matrices, then $\mathbf{GT}$ is a generator matrix in quasi-circulant form of the code $\mathcal{G}_k(\mathbf{g})\mathbf{T}$

The proposition implies that

$$\mathbf{GT} = \begin{pmatrix} \mathbf{G'}_{11} & \cdots & \mathbf{G'}_{1u} \\ \vdots & \ddots & \vdots \\ \mathbf{G'}_{u1} & \cdots & \mathbf{G'}_{uu} \end{pmatrix}$$

where $\mathbf{G'}_{s\ell}$ are submatrices of circulant matrices. This in turn implies that the matrix can be regenerated by storing the u^2 rows of size $\mu \log_2(q^m)$ bits which are the first rows of the matrices $\mathbf{G'}_{s\ell}$.

Corollary 2. To generate the matrix $\mathbf{GT}$ one requires $um^2 \log_2(q)$ bits

Without this particular structure, considering the matrix under systematic form would require $mk(m-k)$ bits. In this sense the gain is significant.

4. STRUCTURE OF THE ENCRYPTION SCHEME

Now we design an encryption scheme similar to that in Loidreau (2017), except for the generation of the public-key. The only difference lies in the public-key generation which is a generator matrix under circulant form of the distorted Gabidulin code. First a designer has to select parameters $q, \mu, \kappa, m = \mu u, k = \kappa u$ and the parameter λ in accordance to the security target.

`KeyGen()`

(1) Pick up randomly a vector $\mathbf{g} \in \mathbb{F}_{q^m}^m$ as in equation (1).
(2) Construct the generator matrix $\mathbf{G}$ for the Gabidulin code $\mathcal{G}_k(\mathbf{g})$ under the form (2).
(3) Pick $\mathcal{V} \subset \mathbb{F}_{q^m}$ a randomly chosen λ-dimensional $\mathbb{F}_q$-vector subspace of $\mathbb{F}_{q^m}$.
(4) Construct a matrix $\mathbf{P} = (\mathbf{P}_{s\ell}) \in GL_m(\mathbb{F}_q)$ where $\mathbf{P}_{s\ell} \in \mathcal{V}^{\mu \times \mu}$, are circulant $\mu \times \mu$ matrices with components in $\mathcal{V}$.
(5) **return** $\mathbf{G}_{pub} = \mathbf{GP}^{-1}$, and $\mathbf{sk} = (\mathbf{G}, \mathbf{P})$

sk stands for the secret key and $\mathbf{G}_{pub}$ is a matrix which generates the public-code $\mathcal{G}_k(\mathbf{g})$. Now Suppose that $\mathbf{p} \in \mathbb{F}_{q^m}^k$ is the plaintext, the encryption procedure is

`Encrypt(p,G`$_{pub}$`)`

(1) Pick $\mathbf{e} \in \mathbb{F}_{2^m}^n$ such that $\mathrm{Rk}(\mathbf{e}) \leq \lfloor (n-k)/2\lambda \rfloor$
(2) **return** $\mathbf{c} = \mathbf{p} \cdot \mathbf{G}_{pub} + \mathbf{e}$

And the decryption procedure for the ciphertext $\mathbf{c}$.

`Decrypt(c,sk)`

- **return** `Decode`$(\mathbf{c} \cdot \mathbf{P}, \mathbf{G})$

where `Decode`$(*, \mathbf{G})$ stands for a decoding algorithm for a Gabidulin code with generator matrix $\mathbf{G}$.

Note that from our construction $\mathbf{G}_{pub} = (\mathbf{G'}_{s\ell})$, where $\mathbf{G'}_{s\ell}$ are submatrices of size $\kappa \times \mu$ of $\mu \times \mu$ circulant matrices. Therefore, to store the public-key, it suffices to store the first rows of each matrix $\mathbf{G'}_{s\ell}$.

5. SECURITY OF THE ENCRYPTION SCHEME

The security of the scheme is related to the difficulty of solving the two following problems

(1) Distinguish the public-code $\mathcal{C}_{pub} = \langle \mathbf{G}_{pub} \rangle$ from a random code
(2) Solve the Rank Bounded Distance Decoding problem of a randomly generated code for the parameters given in the design of the encryption scheme. That is decode errors of rank $\lfloor (n-k)/(2\lambda) \rfloor$ in a k-dimensional code of length n with components in $\mathbb{F}_{2^m}$

Compared to the original scheme, our design has an impact on the first item only. Namely, our construction does not select a particular structured code, but only a particular generator matrix. It can be shown that the security rely completely on the fact that $\mathcal{V}$ remains secret.

However, our proposal introduces a new way to *attack* the first problem which needs careful investigation. Without going into too many details idea of the attack relies on the following considerations

(1) By construction we know that $\mathbf{g} \in \mathcal{V}_{g_1} \times \cdots \times \mathcal{V}_{g_u}$.
(2) From the knowledge of $\mathbf{G}_{pub}$ and of any vector in $\mathbf{g'} = (g'_1, \ldots, g'_u) \in \mathbb{F}_{q^m}^{\mu u}$ such that
$$\mathcal{V}_{g'_1} \times \cdots \times \mathcal{V}_{g'_u} = \mathcal{V}_{g_1} \times \cdots \times \mathcal{V}_{g_u},$$
an attacker can construct a decoder for the public-code thus recovering the plaintext.

A trivial way to achieve this goal is to enumerate vectors $\mathbf{g'} = (g'_1, \ldots, g'_u) \in \mathbb{F}_{q^m}^{\mu u}$, until
$$\forall s = 1, \ldots, u, \ \mathcal{V}_{g'_s} = \mathcal{V}_{g_s}.$$
To obtain a simpler analysis, we relax the condition and say that the attack succeds if

$$\forall s = 1, \ldots, u, \ \begin{cases} \mathcal{V}_{g'_s} \subset \mathcal{V}_{g_s} \\ \mathcal{V}_{g'_s} \neq \{0\} \end{cases}$$

Since for any $s = 1, \ldots, \ell$, the vector space $\mathcal{V}_{g_s}$ is closed under the action of $[u]$. This is equivalent to finding $g'_s \in \mathbb{F}_{q^m} \setminus \{0\}$ such that

$$\forall s = 1, \ldots, u, \ g'_s \in \mathcal{V}_{g_s}$$

Now we can establish the following proposition

Proposition 3. Let $\mathbf{g}' = (g'_1, \ldots, g'_u) \in \mathbb{F}_{q^m}^m$ be randomly and uniformly chosen, then

$$\Pr(\forall s = 1, \ldots, u, \ g'_s \in \mathcal{V}_{g_s}) \approx q^{-\frac{(u-1)m}{2}}$$

6. CONCLUSION

In this abstract, we presented a procedure to reduce significantly the public-key size of a distorted Gabidulin code based encryption scheme. It is of importance that different types of attacks be thouroughly considered before claiming security. However, were this approach sound, then this would enable to provide algebraic code-based public-key cryptosystems with public-key size of a few kilo-bytes.

REFERENCES

Bardet, M., Bros, M., Cabarcas, D., Gaborit, P., Perlner, R.A., Smith-Tone, D., Tillich, J., and Verbel, J.A. (2020). Improvements of algebraic attacks for solving the rank decoding and minrank problems. In *ASIACRYPT 2020*, volume 12491 of *LNCS*, 507–536. Springer.

Delsarte, P. (1978). Bilinear forms over a finite field, with applications to coding theory. *J. Comb. Theory, Ser. A*, 25(3), 226–241.

Gabidulin, E.M. (1985). Theory of codes with maximum rank distance. *Probl. Inf. Transm.*, 21(1), 3–16.

Gabidulin, E.M., Paramonov, A.V., and Tretjakov, O.V. (1991). Ideals over a non-commutative ring and their applications to cryptography. In *Advances in Cryptology - EUROCRYPT'91*, number 547 in Lecture Notes in Comput. Sci., 482–489. Brighton.

Gaborit, P. and Zémor, G. (2015). On the hardness of the decoding and the minimum distance problems for rank codes. *IEEE Trans. Inform. Theory*, 62(12), 7245–7252.

Loidreau, P. (2017). A new rank metric codes based encryption scheme. In *PQCrypto 2017*, volume 10346 of *Lecture Notes in Computer Science*, 3–17. Springer.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

New Results in Multiobjective Model Predictive Control

Gabriele Eichfelder *
Lars Grüne, Lisa Krügel, Jonas Schießl **

* *Institute of Mathematics
Technische Universität Ilmenau, Germany,
(email: gabriele.eichfelder@tu-ilmenau.de)*
** *Chair of Applied Mathematics, Mathematical Institute
Universität Bayreuth, Germany,
(email: lars.gruene, lisa.kruegel, jonas.schiessl@uni-bayreuth.de)*

Keywords: Multiobjective Model Predictive Control, Multiobjective Optimal Control

1. INTRODUCTION

In model predictive control, it is a natural idea that not only one but multiple objectives have to be optimized. This leads to the formulation of a multiobjective optimal control problem (MO OCP). In this talk we introduce a multiobjective MPC algorithm, which yields on the one hand performance estimates for all considered objective functions and on the other hand stability results of the closed-loop solution. To this end, we build on the results in Zavala and Flores-Tlacuahuac (2012); Grüne and Stieler (2019) and introduce a simplified version of the algorithm presented in Grüne and Stieler (2019). Compared to Grüne and Stieler (2019), we allow for more general MO OCPs than in Grüne and Stieler (2019) and get rid of restrictive assumption on the existence of stabilizing stage and terminal costs in all cost components. Compared to Zavala and Flores-Tlacuahuac (2012), we obtain rigorous performance estimate for the MPC closed loop.

2. SETTING AND PRELIMINARIES

For a discrete time nonlinear sytem $x(k+1) = f(x(k), u(k))$, $x(0) = x_0$ with continuous $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, we impose nonempty state and input constraint sets $\mathbb{X} \subseteq \mathbb{R}^n$ and $\mathbb{U} \subset \mathbb{R}^m$, respectively, as well as a nonempty terminal constraint set $\mathbb{X}_0 \subseteq \mathbb{R}^n$, and the set of admissible control sequences for $x_0 \in \mathbb{X}$ up to time $N \in \mathbb{N}$ by $\mathbb{U}^N(x_0) := \{\mathbf{u} \in \mathbb{U}^N \,|\, x_\mathbf{u}(k, x_0) \in \mathbb{X} \;\forall\; k = 1, \dots, N-1 \text{ and } x_\mathbf{u}(N, x_0) \in \mathbb{X}_0\}$.

For given stage costs $\ell_i : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$, $i \in \{1, \dots, s\} =: \mathcal{I}$, $s \in \mathbb{N}_{\geq 2}$, and terminal cost $F_1 : \mathbb{X}_0 \to \mathbb{R}_{\geq 0}$ we define the first cost functional $J_1^N : \mathbb{X} \times \mathbb{U}^N \to \mathbb{R}$ by $J_1^N(x_0, \mathbf{u}) := \sum_{k=0}^{N-1} \ell_1(x_\mathbf{u}(k, x_0), u(k)) + F_1(x(N, x_0))$, and for $i \in \{2, \dots, s\}$, we define cost functionals $J_i^N : \mathbb{X} \times \mathbb{U}^N \to \mathbb{R}$ by $J_i^N(x_0, \mathbf{u}) := \sum_{k=0}^{N-1} \ell_i(x(k, x_0), u(k))$ for horizon $N \in \mathbb{N}_{\geq 2}$. Here, F_1 is defined on the terminal constraint set $\mathbb{X}_0 \subseteq \mathbb{X}$ and we need to ensure that $x(N, x_0) \in \mathbb{X}_0$ by imposing suitable terminal constraints.

⋆ The authors are supported by DFG Grant Gr 1569/13-2.

We consider multiobjective optimal control problems with terminal conditions of the form

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{U}^N(x_0)} \; & J^N(x_0, \mathbf{u}) := \big(J_1^N(x_0, \mathbf{u}), \dots, J_s^N(x_0, \mathbf{u})\big) \\ & x(k+1) = f(x(k), u(k)), \quad k = 0, \dots, N-1 \\ & x(0) = x_0, \quad x(k) \in \mathbb{X} \\ & x(N, x_0) \in \mathbb{X}_0. \end{aligned} \tag{MO OCP}$$

Since we only consider multiobjective optimal control problems with terminal constraint and the terminal constraint $x(N, x_0) \in \mathbb{X}_0$ can generally not be satisfied by all initial values $x_0 \in \mathbb{X}$, we define the feasible set $\mathbb{X}_N := \{x_0 \in \mathbb{X} \mid \exists\; \mathbf{u} \in \mathbb{U}^N : x_\mathbf{u}(k, x_0) \in \mathbb{X},\; k = 1, \dots, N-1,\; x(N, x_0) \in \mathbb{X}_0\} \neq \emptyset$. Further, a pair $(x^e, u^e) \in \mathbb{X} \times \mathbb{U}$ is called equilibrium if $x^e = f(x^e, u^e)$ holds.

In the context of multiobjective optimization we need an appropriate notion of optimality namely the formalization of efficient points and nondominated sets, see, for instance, Ehrgott (2005). Here we denote the set of all efficient solutions of length N for initial value $x_0 \in \mathbb{X}$ by $\mathbb{U}_\mathcal{P}^N(x_0)$. Further, we use the notion of strict dissipativity and some stability results from Grüne and Pannek (2017).

3. A NEW MULTIOBJECTIVE MPC ALGORITHM

In this talk, we combine multiobjective optimization and model predictive control. Thus, we use the MPC theory to solve a multiobjective optimal control problem. Since in the multiobjective case there are several "optimal" (efficient) solutions we have to adapt the "standard" MPC. To this end, we analyze the case of (MO OCP), building on the results in Stieler (2018); Grüne and Stieler (2019), particularly on Algorithm 2 from this last reference. For some theoretical results such as trajectory convergence, performance estimates and stability results, we can use a simplified version of this algorithm, in which we allow for more general MO OCPs. More precisely, we only require that the first stage cost ℓ_1 is strictly dissipative and has a Lyapunov function terminal cost and a corresponding local feedback $\kappa : \mathbb{X}_0 \to \mathbb{U}$, since the first stage cost determine the closed-loop behavior. In contrast we do not require the disappearance or optimality of the other stage costs at

the equilibrium and, thus, the other cost criteria and the selection of the efficient solution determine the transient behavior.

This simplified version is Variant A of the following algorithm, whereas Variant B is Algorithm 2 from Grüne and Stieler (2019).

Algorithm 1 MO MPC with terminal conditions

for $k = 0, \dots, K$:

0. If $k = 0$, set $x(0) = x_0$ and choose an efficient solution $\mathbf{u}^\star_{x(0)} \in \mathbb{U}^N_{\mathcal{P}}(x(0))$ of (MO OCP). Go to 2.
1. **Variant A:**
   If $k \geq 1$, choose a efficient solution $\mathbf{u}^\star_{x(k)}$ of (MO OCP) with $x_0 = x(k)$ so that the inequality
   $$J_1^N(x(k), \mathbf{u}^\star_{x(k)}) \leq J_1^N(x(k), \mathbf{u}_{x(k)}) \tag{1}$$
   holds.
   Variant B:
   If $k \geq 0$, choose a efficient solution $\mathbf{u}^\star_{x(k)}$ of (MO OCP) with $x_0 = x(k)$ so that the inequalities
   $$J_i^N(x(k), \mathbf{u}^\star_{x(k)}) \leq J_i^N(x(k), \mathbf{u}_{x(k)}) \quad \forall i \in \mathcal{I} \tag{2}$$
   hold.
2. For $x := x_{u^*_{x(k)}}(N, x(k))$ set
   $$\mathbf{u}_{x(k+1)} := \left(u^*_{x(k)}(1), \dots, u^*_{x(k)}(N-1), \kappa(x)\right).$$
3. Apply the feedback $\mu^N(k, x(k)) := u^*_{x(k)}(0)$, i.e., evaluate $x(k+1) = f(x(k), \mu^N(k, x(k)))$, set $k = k+1$ and go to 1.

4. RESULTS

For Variant A of Algorithm 1, under the assumption of strict dissipativity of the first stage cost ℓ_1 at an equilibrium (x^e, u^e) for all $(x, u) \in \mathbb{X} \times \mathbb{U}$, we can prove feasibility, trajectory convergence to the equilibrium and a performance estimate for the first cost criterion J_1^N analogously to Stieler (2018); Grüne and Stieler (2019). More precisely, we obtain a bound on the infinite-horizon closed-loop performance of the first cost function of the form
$$J_1^\infty(x_0, \mu^N) \leq J_1^N(x_0, \mathbf{u}^\star_{x_0}). \tag{3}$$
Inequality (3) gives an a-priori bound on the performance only in dependence of the MPC-horizon N, the initial value x_0 and the efficient solution $\mathbf{u}^\star_{x_0}$ chosen in step (0) of Algorithm 1. Further, by using the continuity of the stage costs ℓ_i, $i \in \mathcal{I}$, and the trajectory convergence, we can bound the averaged performance
$$\begin{aligned} \bar{J}_i^\infty(x_0, \mu^N) &:= \limsup_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \ell_i(x(k), \mu^N(x(k))) \\ &\leq \ell_i(x^e, u^e) \end{aligned} \tag{4}$$
for all $i \in \mathcal{I}$ by the corresponding stage cost at the considered equilibrium $\ell_i(x^e, u^e)$.

Moreover, we are interested in stability results. Assuming strict dissipativity of the first stage cost ℓ_1 and some technical inequalities to hold we can prove local asymptotic stability for the MPC closed-loop defined in Algorithm 1, Variant A. To this end, we only use the properties of the first cost function J_1^N and the trajectory convergence. Thus, we can adapt the proof of the single-objective case, see e.g. Grüne and Pannek (2017), and show that the rotated cost function $\widetilde{J}_1^N$ is a time varying Lyapunov function. In our talk we will give insights into the technical assumptions and a sketch of this proof.

Besides the performance of J_1 we are also interested in a non-averaged performance result for J_i^N, $i \in \{2, \dots, s\}$. For this purpose we will use the previous results and combine them with the idea of the performance of single-objective economic MPC without terminal conditions, see Grüne and Pannek (2017). To this end, we consider the trajectories x, which are driven by the efficient solution $\mathbf{u}^\star_{x_0}$. First, we show that the end points of the efficient trajectories are close to the equilibrium because of the stability and the strict dissipative stage cost ℓ_1. Next, in order to establish a performance estimate on J_i, $i \in \{2, \dots, s\}$, we extend the constraint (1) to all $i \in \mathcal{I}$. This way we end up with Algorithm 1 Variant B — an algorithm originally proposed in Grüne and Stieler (2019). However, we still do not require additional properties, such as dissipativity, of the stage cost ℓ_i for $i \geq 2$. Thus, under some technical assumptions we can show that for any $C > 0$ there is a function $\delta \in \mathcal{L}$ such that the performance estimates
$$J_i^K(x_0, \mu^N) \leq J_i^N(x_0, \mathbf{u}^\star_{x_0}) + (K-N)\ell_i(x^e, u^e) + K\delta(N), \tag{5}$$
hold for all $i \in \{2, \dots, s\}$, $N, K \in \mathbb{N}$ with $K \geq N$ and all $x_0 \in \mathbb{X}_N$ with $\|x_0 - x^e\| \leq C$. We remark that for all sufficiently large K the relative error in the estimation above is proportional to $\delta(N)$ and thus decreases to 0 as N tends to infinity. Hence, in terms of the relative error the estimate gives a perfectly useful estimate.

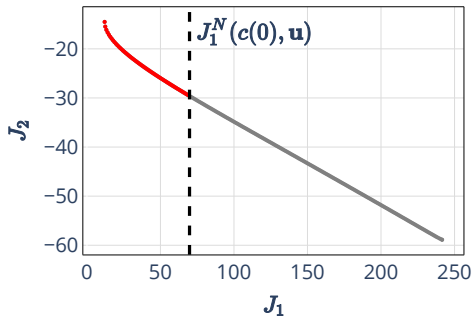
The proofs for all these results will appear in Eichfelder et al. (2022).

5. NUMERICAL SIMULATIONS

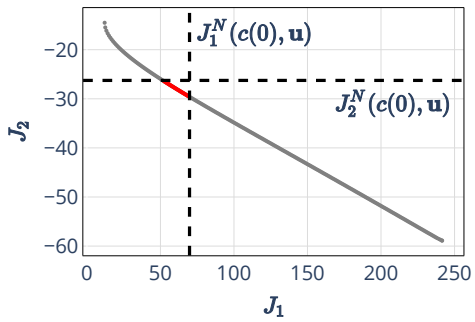
Next, we illustrate the theoretical results of the previous section by numerical simulations. For verifying the theoretical results we use the example of an isothermal chemical reactor, see Diehl et al. (2011); Zavala (2015).

Example 1. We consider a single first-order, irreversible chemical reaction in an isothermal continuous stirred-tank reactor (CSTR). The material balances and the system data are provided in Diehl et al. (2011) and given by
$$\begin{aligned} c_A(k+1) &= c_A(k) + \frac{1}{2}\left(\frac{u(k)}{V}(c_{A_f} - c_A(k)) - k_r c_A(k)\right), \\ c_B(k+1) &= c_B(k) + \frac{1}{2}\left(\frac{u(k)}{V}(c_{B_f} - c_B(k)) + k_r c_B(k)\right), \\ c(0) &= c_0 = (0.4, 0.2) \end{aligned}$$
whereas the stage costs—a tracking type cost forcing the solutions to a desired equilibrium and an economic stage cost maximizing the yield (by minimizing the negative yield) of the reaction—are introduced in Zavala (2015) and given by
$$\begin{aligned} \ell_1(c, u) &= \frac{1}{2}(c_A - \frac{1}{2})^2 + \frac{1}{2}(c_B - \frac{1}{2})^2 + \frac{1}{2}(u - 12)^2, \\ \ell_2(c, u) &= -2uc_B + \frac{1}{2}u. \end{aligned}$$

By setting $\mathbb{X} = [0, 20]$, $\mathbb{U} = [0, 20]$ and $\mathbb{X}_0 = \{(c^e, u^e)\} = \{(\frac{1}{2}, \frac{1}{2}, 12)\}$ we can formulate a multiobjective optimal control problem with terminal constraints of the type (MO OCP). This way, all theoretical assumptions are fulfilled since stabilizing stage cost are a special case of strictly dissipative costs and by setting $\kappa = u^e$ there exists a local feedback with the desired properties. In Figure 1,



Variant A



Variant B

Figure 1. Visualization of step (1) of Algorithm 1

step (1) is visualized for both versions of Algorithm 1 for MPC-horizon $N = 5$. Hence, in Variant A on the left-hand side, only the first cost function J_1^N is restricted whereas on the right-hand side both cost functions are restricted by the additional inequalities (2). Thus, after the first iteration there is still a degree of freedom in choosing the efficient solution. Further, using Variant A of Algorithm 1, we observe trajectory convergence, see Figure 2, and the performance estimate on J_1^N in Figure 3 for MPC-horizon $N = 5$ where the theoretical bound – in dependence of the first chosen efficient solution – is complied.

6. CONCLUSION

We have introduced a new multiobjective MPC algorithm by adapting the algorithm from Grüne and Stieler (2019) and require strong assumptions, namely strict dissipativity and a Lyapunov function terminal cost, only of the first stage cost. For this setting we show that this MOMPC algorithm has the certain desirable properties: feasibility, convergence and stability, and performance results. Finally, we illustrate our theoretical results by means of numerical examples.

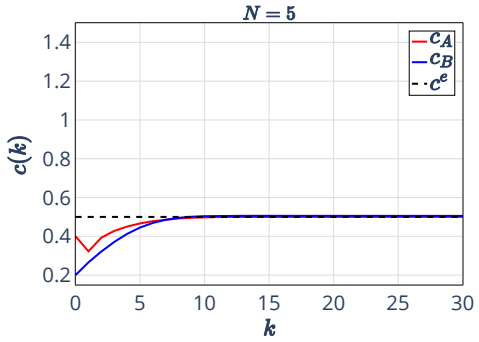


Figure 2. Closed-loop trajectory for $N = 5$

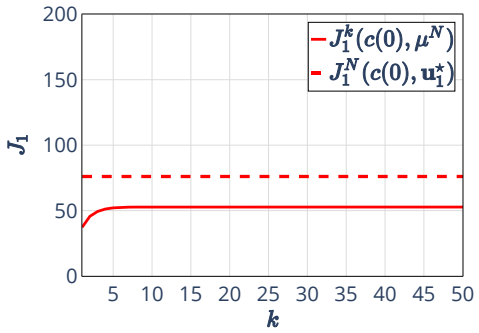


Figure 3. Performance of J_1^N for $N = 5$

REFERENCES

Diehl, M., Amrit, R., and Rawlings, J.B. (2011). A Lyapunov function for economic optimizing model predictive control. *IEEE Transactions on Automatic Control*, 56(3), 703–707. doi:10.1109/TAC.2010.2101291.

Ehrgott, M. (2005). *Multicriteria Optimization*. Springer-Verlag. doi:http://dx.doi.org/10.1007/3-540-27659-9.

Eichfelder, G., Grüne, L., Krügel, L., and Schießl, J. (2022). Relaxed assumptions and a simplified algorithm for multiobjective MPC. In preparation.

Grüne, L. and Pannek, J. (2017). *Nonlinear Model Predictive Control : Theory and Algorithms. 2nd Edition*. Communications and Control Engineering. Springer, Cham, Switzerland.

Grüne, L. and Stieler, M. (2019). Multiobjective model predictive control for stabilizing cost criteria. *Discrete & Continuous Dynamical Systems - B*, 24(8), 3905–3928. doi:https://doi.org/10.3934/dcdsb.2018336.

Stieler, M. (2018). *Performance Estimates for Scalar and Multiobjective Model Predictive Control Schemes*. Ph.D. thesis, Universität Bayreuth, Bayreuth. URL `https://epub.uni-bayreuth.de/3783/`.

Zavala, V.M. (2015). A multiobjective optimization perspective on the stability of economic MPC. *IFAC-PapersOnLine*, 48(8), 974–980. doi: https://doi.org/10.1016/j.ifacol.2015.09.096.

Zavala, V.M. and Flores-Tlacuahuac, A. (2012). Stability of multiobjective predictive control: A utopia-tracking approach. *Automatica*, 48(10), 2627–2632. doi: https://doi.org/10.1016/j.automatica.2012.06.066.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Systems theoretic properties of linear RLC circuits

Timo Reis

Institut für Mathematik, Technische Universität Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (e-mail: timo.reis@tu-ilmenau.de).

Abstract: We consider the differential-algebraic systems obtained by modified nodal analysis of linear RLC circuits from a systems theoretic viewpoint. We derive expressions for the set of consistent initial values and show that the properties of controllability at infinity and impulse controllability do not depend on parameter values but rather on the interconnection structure of the circuit. We further present circuit topological criteria for behavioral stabilizability.
This extended abstract is a shortened version of the full paper Glazov and Reis (2020) which has been accepted for the cancelled MTNS 2020 in Cambridge.

Keywords: electrical circuits, stabilizability, system space, consistent initial values, controllability at infinity, impulse controllability

1. INTRODUCTION

Modified nodal analysis (MNA) is a widely used technique for modelling RLC circuits. It has been first introduced in Ho et al. (1975). It is based on regarding a circuit as a graph, and results in a differential-algebraic model. This model provides a structure which allows a mathematically elegant analysis of essential properties and their physical interpretation. Among these properties is the *index*, i.e., the order of smoothness of perturbations entering the solution of the differential-algebraic equation, see Lamour et al. (2013); Kunkel and Mehrmann (2006); it is shown in Estévez Schwarz and Tischendorf (2000); Günther and Feldmann (1999a,b) that the index is not dependent on system parameters (such as values of resistances, capacitances and inductances), but rather on the interconnection structure, i.e., the topology, of the circuit. Further important possible properties of the circuit system are *stability* and *asymptotic stability*. Whereas MNA models of RLC circuits are always stable as long as the parameter values of resistances, capacitances and inductances are positive, asymptotic stability requires some further conditions. It is shown in Riaza and Tischendorf (2010, 2007); Riaza (2006) that asymptotical stability is guaranteed, if certain parameter-independent criteria on the circuit interconnection structure are fulfilled. The general idea of these articles is used in Berger and Reis (2014), where topological criteria for asymptotic stability and autonomy of the zero dynamics are presented for the purpose of adaptive tracking control of circuits.

In this article, we analyse further systems theoretic properties of the MNA equations. Besides presenting sufficient topological criteria for behavioral stabilizability, we derive the space of consistent initial values, and conclude topological conditions for controllability at infinity and impulse controllability. The key ingredient are the results (Estévez Schwarz and Tischendorf, 2000, Lem. 2.1&2.3) & (Riaza and Tischendorf, 2007, Prop. 4.4&4.5), which give direct links between graph theoretical properties of the circuit to linear algebraic conditions on the involved incidence matrices.

2. CIRCUIT EQUATIONS

The MNA of a linear RLC circuit is given by

$$\tfrac{\mathrm{d}}{\mathrm{d}t}Ex(t) = Ax(t) + Bu(t) \tag{1}$$

with state being composed of vertex potentials, inductive currents, and currents through voltage sources, i.e., $x = (\eta^\top i_\mathcal{L}^\top i_\mathcal{V}^\top)^\top$ and input consisting of voltages at voltage sources and currents at current sources, i.e., $u = (v_\mathcal{V}^\top i_I^\top)^\top$. The matrices E, A, B in (1) are given by

$$sE - A = \begin{bmatrix} sA_\mathcal{C}\mathcal{C}A_\mathcal{C}^\top + A_\mathcal{R}\mathcal{G}A_\mathcal{R}^\top & A_\mathcal{L} & A_\mathcal{V} \\ -A_\mathcal{L}^\top & s\mathcal{L} & 0 \\ -A_\mathcal{V}^\top & 0 & 0 \end{bmatrix},\ B = \begin{bmatrix} -A_I & 0 \\ 0 & 0 \\ 0 & -I_{n_\mathcal{V}} \end{bmatrix}, \tag{2}$$

where s has to be regarded as a formal variable. The expression $sE - A$ is called a *matrix pencil*. Here, $\mathcal{G} \in \mathbb{R}^{n_\mathcal{G}\times n_\mathcal{G}}$, $\mathcal{L} \in \mathbb{R}^{n_\mathcal{L}\times n_\mathcal{L}}$, $\mathcal{C} \in \mathbb{R}^{n_\mathcal{C}\times n_\mathcal{C}}$ are the conductance, inductance and capacitance matrix, and

$$A_\mathcal{R} \in \mathbb{R}^{n_e\times n_\mathcal{R}}, \quad A_\mathcal{L} \in \mathbb{R}^{n_e\times n_\mathcal{L}}, \quad A_\mathcal{C} \in \mathbb{R}^{n_e\times n_\mathcal{C}},$$
$$A_\mathcal{V} \in \mathbb{R}^{n_e\times n_\mathcal{V}}, \quad A_I \in \mathbb{R}^{n_e\times n_I}$$

are the element-specific incidence matrices with sizes $n = n_e + n_\mathcal{L} + n_\mathcal{V}$, $m = n_I + n_\mathcal{V}$. The matrices $\mathcal{G}$, $\mathcal{L}$, $\mathcal{C}$ contain the parameters of capacitances, resistances, and inductances. Further, $A_\mathcal{R}$ is an incidence matrix of the spanning subgraph consisting of all edges that contain resistances. Similarly, the incidence matrices $A_\mathcal{L}, A_\mathcal{C}, A_\mathcal{V}, A_I$ then resp. correspond to the spanning subgraphs with the edges to inductances, capacitances, voltage and current source. An incidence matrix of the finite and loop-free directed graph modeling the circuit is consequently given by $A = [A_\mathcal{R}\ A_\mathcal{L}\ A_\mathcal{C}\ A_\mathcal{V}\ A_I]$. It is also reasonable to assume

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

that the circuit graph is connected, as any connected component corresponds to a subcircuit which does not physically interact with the remaining components, so one may simply consider the connected components separately. It can be inferred from (Estévez Schwarz and Tischendorf, 2000, Lem. 2.1&2.3) that connectedness of a circuit is equivalent to

$$\operatorname{rk}[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}\ A_I] = n_e. \tag{3a}$$

We consider circuits with *passive* devices. This leads to the assumption that the conductance matrix is dissipative, whereas the inductance and capacitance matrices are positive definite, i.e.,

$$\mathcal{G} + \mathcal{G}^\top > 0, \quad \mathcal{L} = \mathcal{L}^\top > 0, \quad \mathcal{C} = \mathcal{C}^\top > 0. \tag{3b}$$

3. REGULARITY AND STABILITY

We take a closer look at the properties of the properties of the pencil $sE - A$ with matrices as in (2). First we recall some results from Berger and Reis (2014).

Proposition 1. Let $E, A \in \mathbb{R}^{n\times n}$ as in (2) and assume that (3) holds. Then there exist invertible $W, T \in \mathbb{R}^{n\times n}$ with

$$W(sE - A)T = \operatorname{diag}(sI - \tilde{A}, sN - I, 0_{n_0,n_0}), \tag{4}$$

where $n_0 \in \mathbb{N}_0$, N is nilpotent with $N^2 = 0$, and $\tilde{A}$ is a square matrix with the property that all its eigenvalues have nonpositive real part. Further, all eigenvalues of $\tilde{A}$ on the imaginary axis are semi-simple (i.e., their respective geometric and algebraic multiplicities coincide). The pencil $sE - A$ further fulfills

$$\begin{aligned} &\ker_{\mathbb{R}(s)}(sE - A) \\ &= \ker_{\mathbb{R}(s)}[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}]^\top \times \{0\} \times \ker_{\mathbb{R}(s)} A_{\mathcal{V}}, \\ &\operatorname{im}_{\mathbb{R}(s)}(sE - A) \\ &= \operatorname{im}_{\mathbb{R}(s)}[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}] \times \mathbb{R}(s)^{n_{\mathcal{L}}} \times \operatorname{im}_{\mathbb{R}(s)} A_{\mathcal{V}}^\top. \end{aligned} \tag{5}$$

A direct consequence of Prop. 1 is that

$$\begin{aligned} \forall \lambda \in \mathbb{C}_+ :\ &\ker_{\mathbb{C}}(\lambda E - A) \\ &= \ker_{\mathbb{C}}[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}]^\top \times \{0\} \times \ker_{\mathbb{C}} A_{\mathcal{V}}, \\ \forall \lambda \in \mathbb{C}_+ :\ &\operatorname{im}_{\mathbb{C}}(\lambda E - A) \\ &= \operatorname{im}_{\mathbb{C}}[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}] \times \mathbb{R}^{n_{\mathcal{L}}} \times \ker_{\mathbb{C}} A_{\mathcal{V}}^\top. \end{aligned} \tag{6}$$

We further characterize *regularity*, i.e., the invertibility of $sE - A$ in $\mathbb{R}(s)^{n\times n}$. Note that regularity translates to the property of a differential-algebraic equation having a solution for all smooth right hand sides, which is moreover unique by specification of the initial condition, see Kunkel and Mehrmann (2006). (Estévez Schwarz and Tischendorf, 2000, Lem. 2.1&2.3) and Prop. 1 allow to characterize regularity in terms of the circuit topology.

Corollary 2. Let $E, A \in \mathbb{R}^{n\times n}$ as in (2) and assume that (3) holds. Then the pencil $sE - A$ is regular, if and only if, the underlying circuit neither contains $\mathcal{V}$-cycles nor I-cuts; equivalently,

$$\ker[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}]^\top = \{0\} \ \wedge\ \ker A_{\mathcal{V}} = \{0\}.$$

Next we consider *generalized eigenvalues* of $sE - A$. This is a complex number λ with $\operatorname{rk}_{\mathbb{C}} \lambda E - A < \operatorname{rk}_{\mathbb{R}(s)} sE - A$. We see from Prop. 1 that all generalized eigenvalues of $sE - A$ have nonpositive real part. In the following we discuss the possible absence of purely imaginary generalized eigenvalues. The absence of generalized eigenvalues on $\overline{\mathbb{C}_+}$ corresponds to stabilizability of the circuit equation $\frac{\mathrm{d}}{\mathrm{d}t}Ex(t) = Ax(t)$. The latter refers to the properties that for all $x_0 \in \mathbb{R}^n$ such that there exists a solution x of $\frac{\mathrm{d}}{\mathrm{d}t}Ex(t) = Ax(t)$ with $Ex(0) = Ex_0$, there also exists a solution x of $\frac{\mathrm{d}}{\mathrm{d}t}Ex(t) = Ax(t)$ with $Ex(0) = Ex_0$ which vanishes at infinity, see (Berger and Reis, 2013, Sec. 5).

Proposition 3. [(Berger and Reis, 2014, Thm. 4.6)] Let $E, A \in \mathbb{R}^{n\times n}$ as in (2) and assume that (3) holds. Then all generalized eigenvalues of $sE - A$ have negative real part, if at least one of the following two assertions holds:

(i) The circuit neither contains $\mathcal{LV}$-cycles except for $\mathcal{V}$-cycles, nor $\mathcal{LCI}$-cuts except for $\mathcal{L}I$-cuts; equivalently,

$$\begin{aligned} \ker[A_{\mathcal{L}}\ A_{\mathcal{V}}] &= \{0\} \times \ker A_{\mathcal{V}}, \\ \wedge\ \ker[A_{\mathcal{R}}\ A_{\mathcal{V}}]^\top &= \ker[A_{\mathcal{R}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}]^\top. \end{aligned}$$

(ii) The circuit neither contains $\mathcal{C}I$-cuts except for I-cuts, nor $\mathcal{LCV}$-cycles except for $\mathcal{CV}$-cycles; equivalently,

$$\begin{aligned} \ker[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}]^\top &= \ker[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{V}}]^\top, \\ \wedge\ \ker[A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}] &= \{0\} \times \ker[A_{\mathcal{C}}\ A_{\mathcal{V}}]. \end{aligned}$$

Prop. 3 slightly generalizes (Riaza and Tischendorf, 2007, Thm. 5.2), where regularity (i.e., the absence of $\mathcal{V}$-cycles and I-cuts) is presumed. Now we combine Prop. 1 with Prop. 3 to show a condition for $\ker_{\mathbb{C}} \lambda E - A = \{0\}$ for all $\lambda \in \overline{\mathbb{C}_+}$. The latter refers to *asymptotic stability*, i.e., all solutions of $\frac{\mathrm{d}}{\mathrm{d}t}Ex(t) = Ax(t)$ vanish at infinity.

Proposition 4. Let $E, A \in \mathbb{R}^{n\times n}$ as in (2) and assume that (3) holds. Then $\ker_{\mathbb{C}} \lambda E - A = \{0\}$ for all $\lambda \in \overline{\mathbb{C}_+}$, if at least one of the following two assertions holds:

(i) The circuit neither contains $\mathcal{LV}$-cycles, nor $\mathcal{LCI}$-cuts except for $\mathcal{L}I$-cuts which are no I-cuts; equivalently

$$\begin{aligned} \ker[A_{\mathcal{L}}\ A_{\mathcal{V}}] &= \{0\}, \\ \wedge\ \ker[A_{\mathcal{R}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}]^\top &= \ker[A_{\mathcal{R}}\ A_{\mathcal{V}}]^\top, \\ \wedge\ \ker[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}]^\top &= \{0\}. \end{aligned}$$

(ii) The circuit neither contains $\mathcal{C}I$-cuts, nor $\mathcal{LCV}$-cycles except for $\mathcal{CV}$-cycles which are no $\mathcal{V}$-cycles; equivalently,

$$\begin{aligned} \ker[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{V}}]^\top &= \{0\}, \\ \wedge\ \ker[A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}] &= \{0\} \times \ker[A_{\mathcal{C}}\ A_{\mathcal{V}}], \\ \wedge\ \ker A_{\mathcal{V}} &= \{0\}. \end{aligned}$$

4. BEHAVIORAL STABILIZABILITY

Loosely speaking, behavioral stabilizability of a differential-algebraic control system (1) means that x can always be asymptotically steered to zero by a suitable choice of the input u. More precisely, for any $x_0 \in \mathbb{R}^n$ for which there exists a control u such that a solution x of (1) with initial condition $Ex(0) = Ex_0$ exists, there especially exists some control u such that a solution x of (1) with $Ex(0) = Ex_0$ exists which vanishes at infinity. It is proven in (Berger and Reis, 2013, Sec. 5) that this is equivalent to

$$\forall\, \lambda \in \overline{\mathbb{C}_+} : \quad \operatorname{rk}_{\mathbb{C}}[\lambda E - A\ B] = \operatorname{rk}_{\mathbb{C}}[\lambda E - A\ B]. \tag{7}$$

Now consider the circuit model $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in (2) and assume that (3) holds. Then

$$\begin{aligned} & \operatorname{im}_{\mathbb{R}(s)}[sE - A\ B] = \operatorname{im}_{\mathbb{R}(s)}(sE - A) \ + \ \operatorname{im}_{\mathbb{R}(s)} B \\ \overset{\text{Prop.1}}{=} \ & \operatorname{im}_{\mathbb{R}(s)}[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}] \times \mathbb{R}^{n_{\mathcal{L}}} \times \operatorname{im}_{\mathbb{R}(s)} \\ & \quad + \operatorname{im}_{\mathbb{R}(s)} A_I \times \{0\} \times \mathbb{R}^{n_{\mathcal{L}}} \\ = \ & \operatorname{im}_{\mathbb{R}(s)}[A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}\ A_I] \times \mathbb{R}^{n_{\mathcal{L}}} \times \mathbb{R}^{n_{\mathcal{V}}} \overset{\text{(3a)}}{=} \mathbb{R}(s)^n. \end{aligned}$$

Likewise, by using (6), the circuit model (2) with assumption (3) fulfills

$$\forall \lambda \in \overline{\mathbb{C}_+} : \ \operatorname{im}_{\mathbb{C}}[\lambda E - A\ B] = \mathbb{C}^n. \tag{8}$$

As a consequence, the circuit model is behaviorally stabilizable if, and only if, $\operatorname{rk}_{\mathbb{C}}[\imath\omega E - A\ B] = n$ for all $\omega \in \mathbb{R}$. This is used in the following result, where we present sufficient conditions for behavioral stabilizability in terms of the circuit topology.

Proposition 5. Let $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in (2) and assume that (3) holds. Then (1) is behaviorally stabilizable, if at least one of the below two statements holds:

(i) The circuit neither contains $\mathcal{L}$-cycles, nor $\mathcal{LC}$-cuts except for $\mathcal{L}$-cuts; equivalently,

$$\begin{aligned} \ker A_{\mathcal{L}} \ &= \{0\}, \\ \wedge \ \ker [A_{\mathcal{R}}\ A_{\mathcal{C}}\ A_{\mathcal{V}}\ A_I]^\top \ &= \ker [A_{\mathcal{R}}\ A_{\mathcal{V}}\ A_I]^\top . \end{aligned}$$

(ii) The circuit neither contains $\mathcal{C}$-cuts, nor $\mathcal{LC}$-cycles except for $\mathcal{C}$-cycles; equivalently,

$$\begin{aligned} \ker [A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{V}}\ A_I]^\top \ &= \{0\}, \\ \wedge \ \ker [A_{\mathcal{L}}\ A_{\mathcal{C}}] \ &= \{0\} \times \ker A_{\mathcal{C}}. \end{aligned}$$

Remark 1. Behavioral stabilizability is implied by *behavioral controllability*, which is defined by concatenation of trajectories (Polderman and Willems, 1998, Def. 5.2.2), and it is algebraically characterized by

$$\forall \lambda \in \mathbb{C} : \ \operatorname{rk}_{\mathbb{C}} [\lambda E - A\ B] = \operatorname{rk}_{\mathbb{R}(s)} [sE - A\ B],$$

see (Berger and Reis, 2013, Cor. 4.3). Note that this property is parameter dependent, as it can be seen from (Polderman and Willems, 1998, Ex. 5.2.13).

5. SYSTEM SPACE

A useful space to understand differential-algebraic systems is the *system space*, which is the minimal subspace $V \subset \mathbb{R}^{n+m}$ in which all solutions $(x(t)^\top u(t)^\top)^\top$ of (1) evolve pointwisely. This space plays a crucial role, for instance in optimal control and dissipativity analysis of differential-algebraic systems, see Reis and Voigt (2015, 2019).

The main result in this section is an expression for the system space of the MNA equations (2).

Theorem 6. Let $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in (2) and assume that (3) holds. Let $Z_{\mathcal{C}}$ and $Z_{\mathcal{RCVI}}$ be basis matrices of $\ker A_{\mathcal{C}}^\top$ and, resp., $\ker[A_{\mathcal{C}}\ A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{V}}\ A_I]^\top$. Then the system space of (1) is given by

$$\ker \begin{bmatrix} Z_{\mathcal{C}}^\top A_{\mathcal{R}} \mathcal{G} A_{\mathcal{R}}^\top & Z_{\mathcal{C}} A_{\mathcal{L}} & Z_{\mathcal{C}}^\top A_{\mathcal{V}} & Z_{\mathcal{C}}^\top A_I & 0 \\ A_{\mathcal{V}}^\top & 0 & 0 & 0 & -I_{n_{\mathcal{V}}} \\ Z_{\mathcal{RCVI}}^\top A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{L}}^\top & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thm. 6 means that a vector $(x_1^\top\ x_2^\top\ x_3^\top\ u_1^\top\ u_2^\top)^\top$ partitioned according to the blocks in $[A\ B]$ as in (2) is in the system space of (1) if, and only if, it satisfies

$$\begin{aligned} Z_{\mathcal{C}}^\top (A_{\mathcal{R}} \mathcal{G} A_{\mathcal{R}}^\top x_1 + A_{\mathcal{L}} x_2 + A_{\mathcal{V}} x_3 + A_I u_1) &= 0, \\ A_V^\top x_1 - u_2 &= 0, \\ Z_{\mathcal{RCVI}}^\top A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{C}}^\top x_1 &= 0. \end{aligned}$$

6. CONSISTENT INITIAL VALUES AND CONTROLLABILITY AT INFINITY

Here we analyze the space of consistent initial values, which is the space of all $x_0 \in \mathbb{R}^n$ for which there exists some control u for which there is a weakly differentiable solution x of (1) with initial condition $x(0) = x_0$. If this space is the entire $\mathbb{R}^n$, then the system (1) is called *controllable at infinity*. It is proven in (Berger and Reis, 2013, Sec. 5) that controllability at infinity is equivalent to $\operatorname{rk}[E\ B] = \operatorname{rk}[E\ A\ B]$. For $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in the circuit model (2) with assumption (3), we can conclude from (8) that $\operatorname{rk}[E\ A\ B] = n$, whence the analysis of controllability at infinity for MNA equations reduces to check whether $\operatorname{rk}[E\ B] = n$. By using $\mathcal{C} > 0$, $\mathcal{L} > 0$, we obtain that $\operatorname{im} E = \operatorname{im} A_{\mathcal{C}} \times \mathbb{R}^{n_{\mathcal{L}}} \times \{0\}$, whence

$$\operatorname{im}[E\ B] = \operatorname{im} A_{\mathcal{C}} \times \mathbb{R}^{n_{\mathcal{L}}} \times \{0\} + \operatorname{im} A_I \times \{0\} \times \mathbb{R}^{n_{\mathcal{V}}}.$$

Controllability at infinity is therefore guaranteed if, and only if, $\operatorname{im}[A_{\mathcal{C}}\ A_I] = \mathbb{R}^{n_e}$ or, equivalently, $\ker[A_{\mathcal{C}}\ A_I]^\top = \{0\}$. We summarize these findings in the following result.

Proposition 7. Let $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in (2) and assume that (3) holds. Then the system (1) is controllable at infinity if, and only if, the underlying circuit does not contain any $\mathcal{RLV}$-cuts; equivalently,

$$\ker [A_{\mathcal{C}}\ A_I]^\top = \{0\}.$$

It can be concluded from (Reis and Voigt, 2019, Lem. 3.7) that the system space $\mathcal{V}_{\text{sys}}$ and the space $\mathcal{V}_{\text{init}}$ of consistent initial values of the system (1) fulfill the identity

$$\mathcal{V}_{\text{init}} = \{x \in \mathbb{R}^n \ : \ \exists u \in \mathbb{R}^m \text{ s.t. } \left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right) \in \mathcal{V}_{\text{sys}}\}. \tag{9}$$

This identity is the essential ingredient in the proof of the following result which contains an expression of the space of consistent initial values for the MNA system.

Theorem 8. Let $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in (2) and assume that (3) holds. Let $Z_{\mathcal{RCVI}}$ and $Z_{\mathcal{CI}}$ be basis matrices of $\ker[A_{\mathcal{C}}\ A_{\mathcal{R}}\ A_{\mathcal{L}}\ A_{\mathcal{V}}\ A_I]^\top$ and, resp., $\ker[A_{\mathcal{C}}\ A_I]^\top$. Then the space of consistent initial values of (1) is given by

$$\ker \begin{bmatrix} Z_{\mathcal{RCVI}}^\top A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{L}}^\top & 0 & 0 \\ Z_{\mathcal{CI}}^\top A_{\mathcal{R}} \mathcal{G} A_{\mathcal{R}}^\top & Z_{\mathcal{CI}}^\top A_{\mathcal{L}} & Z_{\mathcal{CI}}^\top A_{\mathcal{V}} \end{bmatrix}.$$

In the case where there are no $\mathcal{RLV}$-cuts, we can conclude from (Estévez Schwarz and Tischendorf, 2000, Lem. 2.1&2.3) that both Z_{IC} and $Z_{\mathcal{RCVI}}$ are trivial, i.e., these matrices have zero columns. Consequently, we also obtain from Thm. 8 that the absence of $\mathcal{RLV}$-cuts causes that any vector in $\mathbb{R}^n$ is a consistent initial value for the MNA system (cf. Prop. 7).

7. CONSISTENT INITIAL DIFFERENTIAL VALUES AND IMPULSE CONTROLLABILITY

We now consider another type of initialization, namely (1) with initial condition $Ex(0) = Ex_0$. $x_0 \in \mathbb{R}^n$ is called a *consistent initial differential value*, if there exists a control u for which a solution x of (1) with initial condition $Ex(0) = Ex_0$ exists. If this space equals to $\mathbb{R}^n$,

then the system (1) is called *impulse controllable*. It is proven in (Berger and Reis, 2013, Sec. 5) that impulse controllability is equivalent to $\operatorname{rk}[E\ AZ\ B] = \operatorname{rk}[E\ A\ B]$ for some (and hence any) basis matrix Z of $\ker E$. By again using that the circuit model (2) with assumption (3) has the property $\operatorname{rk}[E\ A\ B] = n$, it is impulse controllable if, and only if, $\operatorname{rk}[E\ AZ\ B] = n$. By using that $\mathcal{C} > 0$ and $\mathcal{L} > 0$ by (3b), we obtain that a basis matrix of $\ker E$ is given by $Z = \operatorname{diag}(Z_C, 0, I)$, where Z_C is a basis matrix of $\ker A_C^\top$. Then

$$\begin{aligned}\operatorname{rk}[E\ AZ\ B] &= \operatorname{rk}\begin{bmatrix} A_C\mathcal{C}A_C^\top & 0 & A_\mathcal{R}\mathcal{G}A_C^\top Z_C & A_\mathcal{V} & A_I & 0 \\ 0 & \mathcal{L} & -A_\mathcal{L}^\top Z_C & 0 & 0 & 0 \\ 0 & 0 & -A_\mathcal{V}^\top Z_C & 0 & 0 & I_{n_\mathcal{V}} \end{bmatrix} \\ &= \operatorname{rk}[A_C\ A_\mathcal{R}\mathcal{G}A_\mathcal{R}^\top Z_C\ A_\mathcal{V}\ A_I] + n_\mathcal{L} + n_\mathcal{V}. \end{aligned} \tag{10}$$

If $\ker[A_C\ A_\mathcal{R}\ A_\mathcal{V}\ A_I]^\top \neq \{0\}$, (10) implies $\operatorname{rk}[E\ AZ\ B] < n$. Conversely, if $\ker[A_C\ A_\mathcal{R}\ A_\mathcal{V}\ A_I]^\top = \{0\}$ and $x_1 \in \ker[A_C\ A_\mathcal{R}\mathcal{G}A_\mathcal{R}^\top Z_C\ A_\mathcal{V}\ A_I]^\top$, then $x_1 \in \ker A_C$, i.e., $x_1 = Z_C z_1$ for a vector z_1, and thus $Z_C^\top A_\mathcal{R}\mathcal{G}A_\mathcal{R}^\top Z_C z_1 = 0$. Then $\mathcal{G} + \mathcal{G} > 0$ leads to $A_\mathcal{R}^\top x_1 = A_\mathcal{R}^\top Z_C z_1 = 0$, whence $x_1 \in \ker[A_C\ A_\mathcal{R}\ A_\mathcal{V}\ A_I]^\top = \{0\}$. We summarize these findings in the following result.

Proposition 9. Let $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in (2) and assume that (3) holds. Then the system (1) is impulse controllable if, and only if, the underlying circuit does not contain any $\mathcal{L}$-cuts; equivalently,

$$\ker\left[A_\mathcal{R}\ A_C\ A_\mathcal{V}\ A_I\right]^\top = \{0\}.$$

Theorem 10. Let $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ as in (2) and assume that (3) holds. Let $Z_{\mathcal{RCVI}}$ be a basis matrix of $\ker[A_C\ A_\mathcal{R}\ A_\mathcal{L}\ A_\mathcal{V}\ A_I]^\top$. Then the space of consistent initial differential values of (1) is given by

$$\ker\left[0\ Z_{\mathcal{RCVI}}^\top A_\mathcal{L}\ 0\right].$$

In the case where the circuit does not contain any $\mathcal{L}$-cuts, we can conclude from (Estévez Schwarz and Tischendorf, 2000, Lem. 2.1&2.3) that $Z_{\mathcal{RCVI}}$ are trivial, i.e., it has zero columns. As a consequence, we also obtain from Thm. 10 that in the case of absence of $\mathcal{L}$-cuts, any vector in $\mathbb{R}^n$ is a consistent initial differential value for the MNA system (cf. Prop. 9).

CONCLUSION

We have analyzed the MNA equations of linear time-invariant RLC circuits by using linear algebraic and graph theoretical techniques. Circuit topological criteria for regularity, stability, and behavioral stabilizability, controllability at infinity and impulse controllability have been derived. Moreover, our approach leads to explicit expressions for the system space and for the space of consistent initial (differential) values.

REFERENCES

Berger, T. and Reis, T. (2014). Zero dynamics and funnel control for linear electrical circuits. *J. Franklin Inst.*, 351(11), 5099–5132.

Berger, T. and Reis, T. (2013). Controllability of linear differential-algebraic systems - a survey. In A. Ilchmann and T. Reis (eds.), *Surveys in Differential-Algebraic Equations I*, Differential-Algebraic Equations Forum, 1–61. Springer, Berlin-Heidelberg.

Estévez Schwarz, D. and Tischendorf, C. (2000). Structural analysis for electric circuits and consequences for MNA. *Int. J. Circuit Theory Appl.*, 28(2), 131–162.

Glazov, F. and Reis, T. (2020). Systems theoretic properties of linear rlc circuits. *IFAC-PapersOnLine*, 54(9), 38–45.

Günther, M. and Feldmann, U. (1999a). CAD-based electric-circuit modeling in industry I. Mathematical structure and index of network equations. *Surv. Math. Ind.*, 8, 97–129.

Günther, M. and Feldmann, U. (1999b). CAD-based electric-circuit modeling in industry II. Impact of circuit configurations and parameters. *Surv. Math. Ind.*, 8, 131–157.

Ho, C.W., Ruehli, A., and Brennan, P. (1975). The modified nodal approach to network analysis. *IEEE Trans. Circuits Syst.*, CAS-22(6), 504–509.

Kunkel, P. and Mehrmann, V. (2006). *Differential-Algebraic Equations. Analysis and Numerical Solution.* EMS Publishing House, Zürich, Switzerland.

Lamour, R., März, R., and Tischendorf, C. (2013). *Differential Algebraic Equations: A Projector Based Analysis*, volume 1 of *Differential-Algebraic Equations Forum.* Springer-Verlag, Heidelberg-Berlin.

Polderman, J.W. and Willems, J.C. (1998). *Introduction to Mathematical Systems Theory. A Behavioral Approach.* Springer, New York.

Reis, T. and Voigt, M. (2015). The Kalman-Yakubovich-Popov inequality for differential-algebraic systems: Existence of nonpositive solutions. *Systems Control Lett.*, 86, 1–8.

Reis, T. and Voigt, M. (2019). Linear-quadratic optimal control of differential-algebraic systems: the infinite time horizon problem with zero terminal state. *SIAM J. Control Optim.*, 57(3), 1567–1596.

Riaza, R. (2006). Time-domain properties of reactive dual circuits. *Int. J. Circ. Theor. Appl.*, 34(3), 317–340.

Riaza, R. and Tischendorf, C. (2007). Qualitative features of matrix pencils and DAEs arising in circuit dynamics. *Dynamical Systems*, 22(2), 107–131.

Riaza, R. and Tischendorf, C. (2010). The hyperbolicity problem in electrical circuit theory. *Math. Meth. Appl. Sci.*, 33(17), 2037–2049.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Moment based parametrization of higher order stochastic models

Patrick Dewilde

TUM-IAS, (email: p.dewilde at me.com)

Paper accepted for presentation at MTNS 2020 in Cambridge.
AMS classification: 60G25, 37M05, 30E05

Abstract: The paper reports results on the modeling of related stochastic variables, based on a finite fully ordered sequence of higher order moments (or correlations), and using mutually independent parameters that characterize all solutions that interpolate the given (or measured) data. The results are obtained by determining properties of the hierarchical generalized Hankel matrix of the moments. A system theoretic approach is used to derive the results. It appears that an extension of the related Hamburger-Jacobi orthogonal polynomials to the multivariate case does not suffice to yield a parametrization, but a further reduction of the recursive Cholesky factorization of the moment matrix does.

Keywords: parametrization, multivariate moments, stochastic models

1. INTRODUCTION

Modeling stochastic variables has been a central piece of endeavor in signal processing, leading in particular to the famous Levinson model filters [Kailath (1976)], and their Schur parameter implementation [Dewilde et al. (1978)], in which a (zero means) stochastic process for which n specific covariance data items have been measured is modeled by an artificial linear filter of order n driven by (artificial) white noise. This order n stochastic modeling is known to provide a 'best' possible model when the original process is known to be zero mean Gaussian, where 'best' can be understood as reproducing (interpolating) the measured covariance data and being as unspecified as possible concerning unspecified covariances (i.e., maximum entropy for the unknown parameters, while satisfying the positivity of the covariance matrix). From the model a predictor or estimator can be derived, which, in the specific Gaussian case, is both the conditional expectation as well as the maximum likelihood estimate. It may be remarked that the coefficients of the Levinson polynomials do not form an independent parametrization, but the Schur coefficients do. (The Levinson coefficients have potentially very sensitive internal relations!)

However, many processes are highly non-Gaussian so that non-linear moment information is needed if any modeling accuracy is to be obtained for them [Zarzycki (2004)]. The modeling filter (also to be driven by an artificially constructed process) will have to be non-linear if any accuracy (preferably interpolation of the known data) is to be obtained. Needless to say, the problem just stated is not solvable in all generality, but with the condition of a fully ordered sequence of moments starting at zero it is. This parametrization problem for one variable is fully solved by using the classical Hamburger theory and the parametrization implicit in the related orthogonal polynomial Jacobi recursion [Hamburger (1920-21)]. In this paper, we show how Hamburger's results extend to multiple variables (as they occur in a stochastic process), for as far as independent parametrization is concerned. The general aim is to generate an interpolating pdf of the variables present, i.e., a pdf that is as simple as possible while matching the known data. The present paper expands on the recent publication of the basic theory in [Dewilde (2020)]. In particular, extensions to the nD case as summarized is new.

Notation: I use MATLAB indexing notation, with some abbreviations: $A_{i:j}^{k:l}$ form a (sub)matrix of the (block) matrix A with (block) rows i to j and columns k to l; A' is the transpose of A. A fat index $\underline{\imath}$ indicates a generic order. Expectation is abbreviated by an overline: $\overline{X} = \mathbf{E}X$.

Modeling principle: Given a multidimensional pdf for a joint process involving $n+1$ stochastic variables $Y_0, \cdots, Y_n$, then the best possible model in the pdf sense, is a set of artificial stochastic variables $X_{0:n}$ with the same overall distribution, and the best model for X_0, assuming the $[X_i]_{i=1:n}]$ known, is the conditional probability $\widehat{X}_0 = X_0|_{[X_i]_{i=1:n}]}$. When not the original multidimensional pdf is known, but only some data related to it, then the modeling problem is to find a parametrization of all possible pdf's that match the given data, so that the most adequate solution (for example: the least complex one, or a solution that matches other criteria) can be found. Under *parametrization* we understand a characterization that uses a set of *independent* parameters, i.e., parameters that can be chosen independently from each other and still produce a solution that matches the original data. In our case, the data that we shall match will be a fully ordered set of moments.

The conditional probability is a stochastic variable in one dimension, with a specific pdf for each value of the n-tuple $X_{1:n}$. In the Gaussian case, this pdf is Gaussian as well and its argument is $x - \mu$ with $\mu = -\sum_{i=1:n} a_i x_i$, which is linearly dependent on the $x_{1:n}$ with Levinson coefficients a_i

and a related variance. In the general case the statistics of the estimate is non-Gaussian. The method proposed here determines a discretized version of the pdf, which matches the desired moments exactly. In case not a model is needed, but an estimation or prediction, some strategy is needed to select one representative value. This can be a mean, a median, maximum likelihood or any desirable average.

It has been thought that the moment modeling problem can be solved using an extended version of the classical Schur-Levinson recursion. That is pertinently (but perhaps unfortunately) not true. The modeling problem based on second order covariances in the scalar time invariant case consists in parametrizing positive definite Toeplitz matrices and is essentially different from the moment problem, which consists in characterizing positive definite Hankel matrices. The characteristics, symmetries and number of parameters involved are different, and the resulting models are correspondingly different.

Nonetheless, the solution of both problems uses recursive orthogonalizations of a growing sequence of positive-definite matrices, different in each case. This can already be illustrated on a 2 by 2 positive definite matrix. In the Schur case, three independent numbers $a > 0$, $b > 0$ and $\rho : |\rho < 1|$ parametrize a (non-singular) positive definite matrix as $\begin{bmatrix} a & \rho\sqrt{ab} \\ \rho\sqrt{ab} & b \end{bmatrix}$ while in the moment case (which will be discussed further), the parametrization consists of two positive numbers a, c and an arbitrary number b: $\begin{bmatrix} a & b \\ b & \frac{b^2}{a} + c \end{bmatrix}$ (in both instances, the singular case is a bit more complex.)

The problem to be considered in this paper is the extension of moment parametrization to higher dimensions. Orthogonal polynomials can be extended to the block matrix or multivariable case, as is done for the Levinson case in [Vieira (1977)] and for the moment case in [Arizn-Abarreta and Mañas (2014)]. However, the generalized Jacobi parameters in the multivariate moment case do not amount to a parametrization (nor do Levinson parameters do in the Gaussian estimation case, only Schur parameters are independent [Dewilde and Deprettere (1987)].) A further refinement is needed, and that is the main result of this paper.

The presentation starts out with a brief review of the classical (scalar) Hamburger-Jacobi moment matching case first, but then concentrates on the multivariable stochastic modeling problem given a coherent set of higher order correlation data. Since the matrices to be handled are 'Hankel-like', and Hankel matrices play a key role in system theory, a system-theoretic approach is called for and will be used throughout. Such an approach matches well with the central role played by Cholesky factorization and Schur complementation in the theory.

2. THE 1-D MOMENT MATCHING PROBLEM

Given a required set of moments $\mu_{0:2n}$, can we find a stochastic variable X of minimal complexity that matches them (i.e., $\mu_k = \overline{X^k}$) and find parametrizations for the solutions (notice that, in contrast with the Schur case, one needs $2n$ values for an order n problem)? This question has been very beautifully resolved by Hamburger and Akhiezer, see [Hamburger (1920-21); Akhiezer (1965)], in which the connection between Hankel matrices and Jacobi matrices is exploited to build polynomials that are orthogonal with respect to a positive measure on the real line. Such measures correspond to pdf's (probability density functions) of a stochastic variable. A pdf can only consist of a measurable positive function and positive Dirac impulses. We do not repeat the Akhiezer theory here, but build on it by making the connection with dynamical system theory, in which Hankel matrices play a central role.

A *Jacobi matrix* is by definition a symmetric, tridiagonal matrix, in which the main diagonal is a sequence $\{a_k\}_{k=0:}$ of arbitrary coefficients, and the first off diagonal is a sequence of positive coefficients $\{b_k\}_{k=0:}$, with the termination rule: either all $b_n > 0$ or the series terminate at the first n for which $b_n = 0$. The relation between a positive definite Hankel matrix $H_{\mathbf{n}}$ of dimension $(n+1)^2$ and a related Jacobi matrix can be understood in various ways. Akhiezer uses orthogonal polynomials, with $H_{\mathbf{n}}$ as the Gramian defining the inner product. Here we follow a somewhat different (but of course equivalent) path, based on Cholesky factorization. Let $H_{\mathbf{n}} = L_{\mathbf{n}} L'_{\mathbf{n}}$ be a Cholesky factorization (i.e., $L_{\mathbf{n}}$ is lower triangular).

Case 1: $H_{\mathbf{n}}$ is non-singular. Then $L_{\mathbf{n}}^{-1}$ exists as a lower triangular matrix, and let the rows of $L_{\mathbf{n}}^{-1}$ be defined as $p_{k,:} := [L_{\mathbf{n}}^{-1}]_{k,:}]$. Next we define the k^{th} polynomial $P_k(z) := p_{k,0} + p_{k,1}z + \cdots + p_{k,k}z^k$. It follows immediately that $p_{k,:} H_{\mathbf{n}} p'_{\ell,:} = \delta(k-\ell)$, which is also the usual definition of orthonormality for the $P_k(z)$ with respect to the Gramian $H_{\mathbf{n}}$ (the *inner product* $(P_k(z), P_\ell(z))$ being defined as $p_{k,:} H_{\mathbf{n}} p'_{\ell,:}$). Next, $p_{k,:} H_{\mathbf{n}} e_\ell = 0$ (with $[e_\ell]_j := \delta(\ell - j)$ the ℓ^{th} natural base vector) for $\ell < k$, hence $P_k(z) \perp z^\ell$ for $\ell < k$, and this just because $L_{\mathbf{n}}^{-1} H_{\mathbf{n}} = L_{\mathbf{n}}^{-\prime}$ is upper triangular. It follows that the $P_k(z)$ satisfy a three term recursion

$$zP_k(z) = b_k P_{k-1}(z) + a_k P_k(z) + b_{k+1} P_{k+1}(z) \qquad (1)$$

because 1., both $\{[z^\ell]_{\ell=0:k}\}$ and $\{[P_\ell]_{\ell=0:k}\}$ form bases for the same space and 2., the inner products $(zP_k(z), P_\ell(z)) = (P_k(z), zP_\ell(z))$ are equal, due to the Hankel symmetry.

Infinite non-singular positive definite Hankel matrices can be Cholesky factorized for as much as one wants to go, just by extending the Cholesky factorization of finite restrictions. So, when $H_{\mathbf{m}} = L_{\mathbf{m}} L'_{\mathbf{m}}$ and similarly $H_{\mathbf{n}}$ with $m > n$ are Cholesky factorizations, then $L_{\mathbf{m}}$ is a submatrix of $L_{\mathbf{n}}$, which can be obtained recursively by just extending $H_{\mathbf{m}}$ and $L_{\mathbf{m}}$ recursively (as is usually done in Cholesky factorization). In this way we may consider the infinite Hankel H and its Cholesky factoriztion $H = LL'$, although these infinite matrices are only defined numerically and will typically be very much unbounded. Calculus on such matrices is allowed so long as it remains unilateral, meaning: recursively increasing. This works for the determination of orthonormal polynomials, and the inner products on which they depend, for as long as one works on finitely supported vectors. Let then $[\sigma]_{k,\ell} = \delta(k+1-\ell)$ be the infinitely supported upper shift matrix then the Hankel symmetry on the global Hankel series is

equivalent to $\sigma H = H\sigma'$. Suppose now that $H = LL'$ and L, being lower triangular, is recursively invertible.

Proposition 1. A recursively invertible L is the lower Cholesky factor of a strictly positive definite Hankel matrix iff $J = L^{-1}\sigma L$ is a Jacobi matrix. J characterizes L (and H) uniquely.

Proof One checks that the equations $JL^{-1} = L^{-1}\sigma$ are exactly those that define the three terms recursion. Uniqueness of the solution does the rest.□

The merit of the proposition (and equivalently, the three term recursion) is that it characterizes 'Hankelness' in terms of the Cholesky factor L. The case for a non-singular but finite Hankel matrix is similar, but a bit involved due to the singularity of the finite shift matrix. We shall soon see that it can be derived from a step one extension, which we discuss now.

Case 2: let us assume the positive definite $H_{\mathbf{n}}$ non-singular, and search for a one-step augmentation of $H_{\mathbf{n}}$ to a singular $H_{\mathbf{n+1}} = \begin{bmatrix} H_{\mathbf{n}} & \mu_{n+1:2n+1} \\ \mu'_{n+1:2n+1} & \mu_{2n+2} \end{bmatrix}$. μ_{2n+1} can be chosen arbitrarily, and next $\mu_{2n+2} = \mu'_{n+1:2n+1} H_{\mathbf{n}}^{-1} \mu_{n+1:2n+1}$ must be the Schur complement of μ_{2n+2} if $H_{\mathbf{n+1}}$ is to be singular. From system theory (realizations), we now know that there exist an $(n+1)\times(n+1)$ matrix A and (n+1)-vectors B and C' so that $\mu_k = CA^kB$. Additionally, because $H_{\mathbf{n+1}}$ is symmetric, one can choose A symmetric and $C' = B$. It follows that the eigenvalues of A are all real and there shall exist a unitary state transformation matrix U so that $UAU' = \mathrm{diag}[\alpha_i]$ is diagonal. Redefining the realization $A \leftarrow UAU', B \leftarrow \beta = Ub, C' \leftarrow Uc' = \beta'$ then produces $\mu_k = \sum_{i=0:n} \beta_i^2 \alpha_i^k$. It follows that the moments $\mu_{0:2n+2}$ can be generated by an *atomic* pdf $p(x) = \sum_{i=0:n} \beta_i^2 \delta(x-\alpha_i)$. To this there corresponds a rational *generating function* $G_X(z) = 1 + \mu_1 z + \mu_2 z^2 + \cdots = \sum_{i=1:n} \frac{b_i^2}{1-za_i}$ and we have that $G_X(z)$ matches the $2n+2$ first moments of $G_Y(z)$.

Remark 1: once $H_{\mathbf{n+1}}$ is singular of rank n, then all subsequent H_k $(k > n)$ are singular of rank n as well, and all moments are given by $\mu_k = \beta' A^k \beta$. Every given partial series $\mu_{0:2n}$ with $H_{\mathbf{n}}$ non-singular has a rational generating function of degree n, and can be continued with all subsequent $H_{\mathbf{k}}$, $k > n$, singular.

Remark 2: the generating function $G(z)$ is generally highly unstable, and should be handled using polynomial calculus on unilateral series.

Remark 3: the generating function corresponding to a zero means Gaussian variable of variance σ has $\mu_k = (k-1)!!\sigma^k$, where $(k-1)!!$ equals 0 when k is odd and otherwise $1\cdot 3\cdots(k-1)$. Hence, it converges nowhere in the complex plane (except at zero).

A further observation leads to another interesting result connecting realization theory with Jacobi matrices:

Proposition 2. The realization for $G(z) = \sum_{k=0:} B'A^kBz^k$ based on a Cholesky factorization of the non-singular $H_{\mathbf{n}} = L_{\mathbf{n}} L'_{\mathbf{n}}$ is given by $\{A, B, B'\} \leftarrow \{J_n, e_0, e'_0\}$, in which $[e_0]_k = \delta(k)$ and $J_{\mathbf{n}}$ is the n^{th} order Jacobi matrix related to the singular $H_{\mathbf{n+1}}$. Conversely, any such realization produces a moment series μ_k with non-singular positive $H_{\mathbf{n}}$ and singular $H_{\mathbf{n+1}}$.

Proof (sketch) The Cholesky factorization produces the relation, for $1 < k \le n$,

$$p_k^{\rightarrow} = b_{k-1}p_{k-1} + a_k p_k + b_k p_{k+1}, \tag{2}$$

with $p_k^{\rightarrow}$ a right shift version of p_k. This relation is equivalent to (1).

3. GENERALIZED HANKELS IN TWO AND N VARIABLES

Two stochastic variables (2D). A full moment matrix in two dimensions is $\overline{\mathbf{X}\mathbf{X}'}$ for $\mathbf{X}' := \left[1 \middle| X\ Y \middle| X^2\ XY\ Y^2 \middle| \cdots\right]$. Such matrices (also for more variables) have a *hierarchical* Hankel structure, in the sense that blocks on anti-diagonals contain the same elements and are internally Hankel as well, with compatible ordering structures. Connected to the

$H_3 =$

1	μ_{10}	μ_{01}	μ_{20}	μ_{11}	μ_{02}	μ_{30}	μ_{21}	μ_{12}	μ_{03}
μ_{10}	μ_{20}	μ_{11}	μ_{30}	μ_{21}	μ_{12}	μ_{40}	μ_{31}	μ_{22}	μ_{13}
μ_{01}	μ_{11}	μ_{02}	μ_{21}	μ_{12}	μ_{03}	μ_{31}	μ_{22}	μ_{12}	μ_{04}
μ_{20}	μ_{30}	μ_{21}	μ_{40}	μ_{31}	μ_{22}	μ_{50}	μ_{41}	μ_{32}	μ_{23}
μ_{11}	μ_{21}	μ_{12}	μ_{31}	μ_{22}	μ_{13}	μ_{41}	μ_{32}	μ_{23}	μ_{14}
μ_{02}	μ_{12}	μ_{03}	μ_{22}	μ_{13}	μ_{04}	μ_{32}	μ_{23}	μ_{14}	μ_{05}
μ_{30}	μ_{40}	μ_{31}	μ_{50}	μ_{41}	μ_{32}	μ_{60}	μ_{51}	μ_{42}	μ_{33}
μ_{21}	μ_{31}	μ_{22}	μ_{41}	μ_{32}	μ_{23}	μ_{51}	μ_{42}	μ_{33}	μ_{24}
μ_{12}	μ_{22}	μ_{13}	μ_{32}	μ_{23}	μ_{14}	μ_{42}	μ_{33}	μ_{24}	μ_{15}
μ_{03}	μ_{13}	μ_{04}	μ_{23}	μ_{14}	μ_{05}	μ_{33}	μ_{24}	μ_{15}	μ_{06}

Fig. 1. The 2D Hankel structure in H_3

2D case we can define a 2D generating function, very much in the same way as in the 1D case, but now introducing two variables z_1 and z_2, summed by order and then as $\mu(I-i,i)_{i=0:I} = (\overline{X^{I-i}Y^i})_{i=0:I}$ within order I: $G(z_1,z_2) = \sum_{I=0:,(i=I:0,j=I-i)} \mu_{i,j} z_1^i z_2^j$, which can also be seen as the unilateral expansion in 2D of $\mathbf{E}\frac{1}{(1-z_1X)(1-z_2Y)}$. We now have almost immediately:

Theorem 1. A rational generating function $G(z_1,z_2)$ corresponds to a 2D atomic distribution.

The **proof** is based on considering marginals and applying the 1D theory on them.

Orthogonalization of the non-singular $\mathbf{X}$ series in order produces a 2D orthogonal sequence and corresponding block tri-diagonal Jacobi matrix of the form

$$\begin{bmatrix} P_{00}X \\ P_{00}Y \\ \hline P_{10}X \\ P_{01}X \\ P_{01}Y \\ \hline P_{20}X \\ P_{11}X \\ P_{02}X \\ P_{02}Y \end{bmatrix} = \left[\begin{array}{cc|ccc|ccc|cccc} * & a & & & & & & & & & & \\ * & * & c & & & & & & & & & \\ \hline a & * & h & d & & & & & & & & \\ 0 & h & * & e & f & & & & & & & \\ c & * & * & * & * & g & & & & & & \\ \hline 0 & d & e & * & i & j & * & & & & & \\ 0 & 0 & f & i & * & k & * & * & & & & \\ 0 & 0 & 0 & j & k & * & * & * & * & & & \\ 0 & 0 & g & * & * & * & * & * & * & * & * \end{array}\right] \begin{bmatrix} P_{00} \\ \hline P_{10} \\ P_{01} \\ \hline P_{20} \\ P_{11} \\ P_{02} \\ \hline P_{30} \\ P_{21} \\ P_{12} \\ P_{03} \end{bmatrix} \tag{3}$$

in which the upper block diagonal is non-singular and redundancies are indicated with like literals. Unlike in the 1D case, the orthogonalization does not deliver a full parametrization. The coefficients on the rightmost diagonal can be chosen independently, but not the others, and there are more hidden dependencies than indicated by

the literals. Let us therefore consider a full order update of the 2D hierarchical Hankel matrix, namely $H_{\mathbf{i}} \Rightarrow H_{\mathbf{i+1}}$. $H_{\mathbf{i+1}}$ inherits $H_{\mathbf{i}}$ and $H_{0:i}^{i+1}$ from the previous step, only H_i^{i+1} and H_{i+1}^{i+1} are new. H_i^{i+1} is Hankel and of odd order: its entries may be chosen arbitrarily. H_{i+1}^{i+1} has to be Hankel and such that the whole $H_{\mathbf{i+1}}$ is positive definite, which will be the case iff $H_{i+1}^{i+1} \geq S_{i+1}$, for the *Schur residue* $S_{i+1} := L_{i+1}^{0:i}(L_{i+1}^{0:i})'$ (notice: $L_{i+1}^{0:i}$ is known at this point.) The Schur residue itself is not Hankel, but the following holds:

Proposition 3. A non-singular L_{i+1}^{i+1} can be fully parametrized by arbitrary strictly positive numbers on the main diagonal, and arbitrary numbers on the first lower diagonal.

This solves the 2D parametrization problem for the non-singular case.

The singular case. Any discretized pdf as final result of a parametrization, must involve steps that finally make the full moment matrix singular. This is done by deciding on the maximal degree of one of the variables, making the column corresponding to that variable singular, and deleting all columns corresponding to monomials having that power as a factor. The non-singular proceeding can then be continued on the pruned system. For 1D systems, one can just puts the last Jacobi coefficient $b_n = 0$, and the Hankel remains singular and fully determined from that point on. This produces a (non-uniform) discretization of the pdf. For 2D systems, the progressive order can be reduced in two steps: a first reduction produces a polynomial algebraic relation between X and Y, and a second reduction makes the complete system redundant from that order on, resulting in a 2D discretization of the pdf. These two steps can be combined in one. It turns out that this generalizes to nD systems, which can be reduced in n steps, using linear dependencies and generalized Hankel symmetries. Each time a singularity is introduced in a column corresponding to a monomial, say M, then all rows and columns corresponding to monomials that have M as a factor are determined, and the characterization can be continued on a reduced set of moments.

The nD case: The procedure described for the 2D case generalizes to n variables, with mostly technical complications, due to the increased hierarchy. Keeping the reverse lexicographic order at each degree we now have as global stochastic vector

$$\mathbf{X}' := [1|X, Y, Z, \cdots|X^2, XY, XZ, \cdots]. \tag{4}$$

The basic principles remain: (1) odd orders can be chosen arbitrarily (in the non-singular case, in the singular cases many are determined by singularity reductions), and (2) only the central order update block has to be parametrized further recursively so that it becomes larger than the Schur residue, and this is achieved again by specifying its diagonal and subdiagonal entries, in a recursive way corresponding to a one-order lower update. Fig. 2 shows the order of parameter determination for order 4 in the 3D case, starting form the left upper corner: regular Jacobi for the first block, the next two blocks are odd and free, the next diagonal block requires an application of proposition 3 etc. It follows that recursive applications of filling in odd blocks followed by applications of proposition 3 lead to the solution for the 3D case. Likewise, the nD case reduces to recursive applications of the (n-1)D case etc. Only blocks

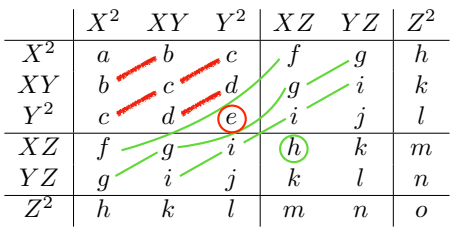


Fig. 2. The order update illustrated on order 4 in the 3D case.

on main diagonals specify free positive entries on top of Schur residues, while odd blocks can be chosen at will at all levels.

The considerations given so far lead to general extensions as well: assuming non-singularity up to some level n, obtained by adding central $[H_{\mathbf{k}}]_{k,k}$'s which are strictly larger that the k^th order Schur complement S_k for $k < n$, when a sufficient level of n of sampling accuracy is obtained, a new $[H_{\mathbf{n}}]_{n,n}$ has to be added for which $[H_{\mathbf{n}}]_{n,n} - S_{\mathbf{n}}$ is rank deficient of order 2. When that is the case, then all subsequent Hankels will be rank deficient, which, after at most n such steps, leads to an atomic distribution that meets the moment data up to order n.

So far for the basic parametrization. The method can be refined considerably by using the connection between Cholesky factors and Jacobi parameters in scalar cases or Jacobi blocks in higher order cases. How this works is beyond the present paper.

REFERENCES

Akhiezer, N. (1965). *The Classical Moment Problem.* Oliver and Boyd, Edinburgh.

Arizn-Abarreta, G. and Mañas, M. (2014). Multivariate orthogonal polynomials and integrable systems. *Advances in Mathematics*, 302, 628–739.

Dewilde, P. and Deprettere, E. (1987). Approximate inversion of positive matrices with application to modeling. In *Modeling, robustness and sensitivity reduction in control systems*, 212–238. Springer, NATO ASI Series.

Dewilde, P., Vieira, A., and Kailath, T. (1978). On a generalized Szegö-Levinson realization algorithm for optimal linear predictors based on a network synthesis approach. *IEEE Trans. Circuits Syst.*, 25(9), 663–675.

Dewilde, P. (2020). Stochastic models based on moment matching. *Communications in Information and Systems*, 20(2), 209–248.

Hamburger, H. (1920-21). Ueber eine erweiterung des stieltjesschen momentproblems. *Math. Ann.*, 81,82, 235–319,120–164.

Kailath, T. (1976). *Lectures on Linear Least-Squares Estimation.* Springer Verlag, CISM Courses and Lectures No. 140, Wien, New York.

Vieira, A. (1977). *Matrix Orthogonal Polynomials, with Applications to Autoregressive Modeling and Ladder Forms.* Ph.D. thesis, Stanford University.

Zarzycki, J. (2004). Multidimensional non-linear schur parametrization of nongaussian stochastic signals. parts i-iii. *Multidimensional Systems and Signal Processing*, 15, 217–241,243–275,313–340.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Learning and numerical integration of Lagrangian dynamics

Christian Offen * **Sina Ober-Blöbaum** *

* *Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany (e-mail: christian.offen@uni-paderborn.de).*

Abstract: Hamilton's principle is one of the most fundamental principle in physics. Incorporating the principle into data-driven models of dynamical systems guarantees that motions share important qualitative properties with the real system, such as energy or momentum conservation. To learn Lagrangian dynamics, we propose to learn inverse modified Lagrangians related to variational integrators instead of attempting to learn an exact Lagrangian, as is typically done in the literature. The key advantage is that inverse modified Lagrangians can be learned from snapshots of position data of observed trajectories directly without approximating velocities or acceleration data. This is beneficial when snapshot times are large. Moreover, when inverse modified Lagrangians are integrated using a variational method, discretisation errors are compensated for. Therefore, large step-sizes can be used while maintaining high accuracy and tiny energy errors.

Keywords: Hamilton's principle, variational integrators, backward error analysis
2020 MSC: 65P10, 93B30, 68T05

1. INTRODUCTION

Physics informed learning describes the incorporation of prior physical knowledge into machine learning based models for unknown dynamical systems to improve accuracy and reliability of predicted motions. One of the most fundamental physical principles is *Hamilton's principle*. It states that a motion $q\colon [t_0, t_1] \to Q$ of a dynamical system connecting two points q_0 and q_1 in a manifold Q extremises an action

$$S(q) = \int_{t_0}^{t_1} L(t, q(t), \dot{q}(t)) \tag{1}$$

for some $L\colon \mathbb{R} \times TQ \to \mathbb{R}$ called *Lagrangian*. From the stationarity condition on S, *Euler–Lagrange equations* $0 = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}}(t,q,\dot{q}) - \frac{\partial L}{\partial q}(t,q,\dot{q})$ can be derived.

A typical strategy to incorporate variational structure into learned models of dynamical systems is to learn the Lagrangian L using neural networks (LNN – see Cranmer et al. (2020)) or Gaussian Processes (LGP – see Ober-Blöbaum and Offen (2022)). In a second step, motions are predicted by integrating the learned dynamical system using a numerical method.

Existing approaches such as LNN assume the availability of trajectory data consisting of position data, velocity data, and acceleration data to learn the Lagrangian L. Refinements such as in Takehiro Aoshima (2021) have been introduced to eliminate some of the data requirements but these assume further prior knowledge about the form of L. In the general case and if only snapshots of positions of trajectories are available, velocity and acceleration data need to be approximated to learn L. We will show that if snapshot times are large, this can cause unacceptably large and biased errors in predicted motions. Further discretisation errors occur when the learned dynamical system is integrated. If step-size adaptive methods are employed, this will come at the cost of not preserving variational structure during integration. Numerical motions then show unphysical behaviour, such as dissipating energy even though the exact system might be energy conserving, which is problematic in long-term simulations or when periodic orbits should be detected. If, on the other hand, variational integrators are used, such that the numerical motions inherit the variational structure and thus much of the qualitative properties of the system (1) (such as energy conservation for autonomous L), error tolerances might require tiny step-sizes.

As a remedy, we introduce *Lagrangian Shadow Integrators* (LSI) which learn an *inverse modified Lagrangian* L_{imd} instead of L. Inverse modified Lagrangians can be learned directly from position data of trajectories such that there is no need to approximate velocity data or acceleration data for the training process. Moreover, inverse modified Lagrangians compensate for discretisation errors in the integration step. In this way, no artificial errors are introduced into the predictions while incorproating variational structure at the same time. We can, therefore, handle snapshot data of trajectories which relate to moderate to large snapshot times. The step-size selection for the integration becomes independent of accuracy requirements, which is beneficial in long-term simulations. Additionally, the Lagrangian L can be computed from L_{imd} using variational backward error analysis.

Extending our framework developed in Ober-Blöbaum and Offen (2022), we will incorporate noise in the derivation of Gaussian Processes based LSI, explain how to handle data with inconsistent snapshot times, provide an analysis employing the generalised midpoint rule, and point out new directions of our research.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

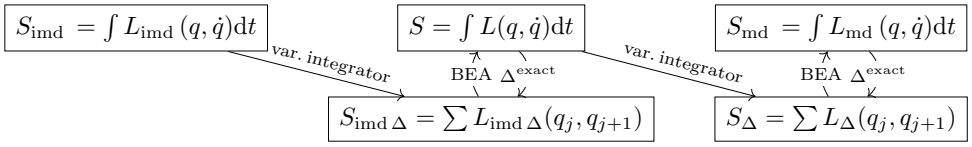


Fig. 1. Illustration of modified Lagrangians and inverse modified Lagrangians as explained in section 2.2. BEA stands for (variational) backward error analysis (Vermeeren (2017)), Δ^{exact} associates exact discrete variational principles.

2. THE LSI FRAMEWORK

2.1 Overview of the LSI procedure

Let us give an overview of the steps involved in Lagrangian Shadow Integration (LSI). We will restrict ourself to autonomous Lagrangians.

- Preparation and learning:
  (1) A variational integrator and a step-size h are selected.
  (2) An inverse modified Lagrangian L_{imd} is learned directly from position data of trajectories.
  (3) A formula for the Lagrangian L is computed from L_{imd} using variational backward error analysis.
- Computation of a motion to initial data $(q_0, \dot{q}_0)$:
  (1) Compute the conjugate momentum at time $t = 0$ as $p_0 = \frac{\partial L}{\partial \dot{q}}(q_0, \dot{q}_0)$.
  (2) Solve $\nabla_1 L_{\mathrm{imd}\,\Delta}(q_0, q_1) + p_0 = 0$ for q_1.
  (3) Solve for q_{j+1} iteratively
  $$\nabla_2 L_{\mathrm{imd}\,\Delta}(q_{j-1}, q_j) + \nabla_1 L_{\mathrm{imd}\,\Delta}(q_j, q_{j+1}) = 0.$$
- Post-processing (optional):
  (1) If velocity data is required at step q_j, solve for $\dot{q}_j$
  $$\nabla_2 L_{\mathrm{imd}\,\Delta}(q_{j-1}, q_j) = \frac{\partial L}{\partial \dot{q}}(q_j, \dot{q}_j). \qquad (2)$$

Above, $L_{\mathrm{imd}\,\Delta}$ denotes the discrete Lagrangian obtained from L_{imd} using the selected variational integrator. The expressions $\nabla_1 L_{\mathrm{imd}\,\Delta}$ and $\nabla_2 L_{\mathrm{imd}\,\Delta}$ denote partial derivatives with respect to the first or second argument of $L_{\mathrm{imd}\,\Delta}$, respectively. In conclusion, the computation of motions of the learned system proceeds like applying the variational integrator to the inverse modified Lagrangian L_{imd}. Whenever velocity data needs to be related to conjugate momenta, the computed formula for L is used. This computed formula can be derived once and for all for each variational integrator.

In the following, we will introduce the notion of inverse modified Lagrangians and provide details on the preparation and learning steps of LSI.

2.2 Modified and inverse modified Lagrangians

If $L\colon TQ \to \mathbb{R}$ is a regular Lagrangian, the *exact discrete Lagrangian* to a step-size h is given as $L_{\Delta}^{\mathrm{exact}}(q_0, q_1) = \int_{t_0}^{t_0+h} L(q(t), \dot{q}(t))\,\mathrm{d}t$, where $q\colon [t_0, t_0 + h] \to Q$ fulfils the Euler–Lagrange equations and the boundary conditions $q(t_0) = q_0$, $q(t_1) = q_1$. $L_{\Delta}^{\mathrm{exact}}(q_0, q_1)$ exists, provided that $q_0, q_1 \in Q$ are sufficiently close to one another.

The notion of inverse modified Lagrangians is illustrated in Fig. 1: let L be a regular Lagrangian. A variational integrator translates a continuous action functional $S(q) = \int_{t_0}^{t_1} L(q(t), \dot{q}(t))$ to a discrete action functional $S_{\Delta}(q) = \sum_{j=0}^{N-1} L_{\Delta}(q_j, q_{j+1})$. A *modified Lagrangian* L_{md} is a Lagrangian such that its exact discrete Lagrangian $L_{\mathrm{md}\,\Delta}^{\mathrm{exact}}$ coincides with L_{Δ}.

Remark 1. Modified Lagrangians L_{md} exist as formal power series in the step-size h and can be computed by *variational backward error analysis (BEA)*, see Vermeeren (2017).

We introduce an inverse version of modified Lagrangians.

Theorem 2. (Ober-Blöbaum and Offen (2022)). For a regular Lagrangian L and a consistent variational integrator there exists a formal power series L_{imd} in h such that the discrete Lagrangian $L_{\mathrm{imd}\,\Delta}$ coincides with the exact discrete Lagrangian $L_{\Delta}^{\mathrm{exact}}$ up to any order in h.

Example 3. The generalised variational midpoint rule assigns to a Lagrangian L the discrete Lagrangian

$$L_{\Delta}(q_0, q_1) = L\left((1-a)q_1 + aq_0, (q_1 - q_0)/h\right) \qquad (3)$$

with a parameter $a \in \mathbb{R}$. The standard midpoint rule uses $a = 1/2$ and is second order accurate. Using variational backward error analysis, we compute the differential operator

$$\Lambda = \mathrm{Id} + \frac{h}{2}(1-2a)\dot{q}\frac{\partial}{\partial q} + \frac{h^2}{24}\Big(\Big(\frac{\partial}{\partial q} - \dot{q}\frac{\partial^2}{\partial q \partial \dot{q}}\Big)^2 \Big/ \frac{\partial^2}{\partial \dot{q}^2} + 2(1 + 6a(a-1))\dot{q}^2\frac{\partial^2}{\partial q^2}\Big) + \mathcal{O}(h^3).$$

The modified Lagrangian is given as $L_{\mathrm{md}} = \Lambda(L)$. If L_{imd} is the inverse modified Lagrangian, then $\Lambda(L_{\mathrm{imd}}) = L$ to all orders in h. A derivation of Λ and Λ^{-1} including higher order terms can be found in the Mathematica script of the accompanying source code. [1]

2.3 Learning of L_{imd}

We now show how to learn L_{imd} from position data of trajectories. In contrast to the theoretical considerations in the previous section, here L_{imd} is *not* a formal power series but a function $L_{\mathrm{imd}} : TQ \to \mathbb{R}$ represented by either an artificial neural network (ANN) or a Gaussian Process and corresponds to a fixed step-size h.

Preparation of training data Before we start learning, we need to make sure that all observed discrete trajectories have consistent snapshot times h_{snap} compatible with h.

Method 1. If the snapshot times of l observed discrete trajectories $(q_j^{(1)})_j, \ldots, (q_j^{(l)})_j$ are in a rational relation such that $h_{\mathrm{snap}} = r_1/m, r_2/m, \ldots, r_l/m$ with (lowest) common denominator m, then the trajectories can be split up such that they all refer to the same snapshot time: if r is the least common multiple of $r_1, \ldots, r_l$, we split the ith trajectory into $(q^{(i)}_{(r/r_i)j})_j$, $(q^{(i)}_{(r/r_i)j+1})_j, \ldots, (q^{(i)}_{(r/r_i)j+(r/r_i)-1})_j$

[1] https://github.com/Christian-Offen/LagrangianShadowIntegration

for $i = 1, \dots, l$. The snapshot times of the new trajectories are $h_{\text{snap}} = r/m$.

Method 2. An alternative to method 1 is to introduce dummy variables for missing observations into the trajectories such that all trajectories have snapshot time $h_{\text{snap}} = 1/m$. The values of the dummy variables will be interpreted as additional parameters in the learning process and will be subject to the optimisation procedure.

Method 3. If all trajectories have snapshot time h_{snap} and $h = kh_{\text{snap}}$ for $k \in \mathbb{N}$, then each trajectory $(q_j)_j$ is split up into k discrete trajectories $(q_{kj})_j$, $(q_{kj+1})_j$, …, $(q_{kj+k-1})_j$ which correspond to a snapshot time of h.

Method 4. As in method 2, if $h = h_{\text{snap}}/k$ for $k \in \mathbb{N}$ we can fill up missing observations with dummy variables which are interpreted as parameters subject to the learning process.

Any rational snapshot time can be obtained with a combination of the above methods. However, methods 1 and 3 cause a loss of training data, since some interrelations of points are forgotten. Methods 2 and 4, however, introduce more parameters that need to be fitted. In the following, we assume that $h = h_{\text{snap}}$.

Training an ANN Let $L_{\text{imd}\,\Delta}$ denote the discrete Lagrangian assigned to L_{imd} by the integrator. For a weighting factor $\theta > 0$, we employ the loss function $\ell = \ell_{\text{data}} + \theta\ell_{\text{normal}}$ with

$$\ell_{\text{data}} = \sum\nolimits_{(q_j)_j \in \text{trj}} \sum\nolimits_{j=2}^{m-1} \|\text{DEL}_{\text{imd}}\,(q_{j-1}, q_j, q_{j+1})\|^2.$$

The sum is taken over all discrete observations of trajectories $(q_j)_j = (q_1, \dots, q_m)$. DE denotes the discrete Euler–Lagrange operator, i.e. $\text{DEL}_{\text{imd}}\,(q_{j-1}, q_j, q_{j+1})$ denotes $\nabla_2 L_{\text{imd}\,\Delta}(q_{j-1}, q_j) + \nabla_1 L_{\text{imd}\,\Delta}(q_j, q_{j+1})$. Minimisation of ℓ_{data} seeks to fulfil the discrete Euler–Lagrange equations (DELs) on all triples (q_{j-1}, q_j, q_{j+1}) in the trajectory data. To avoid learning a constant L_{imd} (whose DELs are trivial), we consider the normalisation

$$\ell_{\text{normal}} = \Big| c - \frac{1}{2^{2n}} \sum\nolimits_{\text{corners of } E^{2n}} \frac{\partial L_{\text{imd}}}{\partial \dot{q}^1} \cdot \ldots \cdot \frac{\partial L_{\text{imd}}}{\partial \dot{q}^n} \Big|^2$$

for $c \in \mathbb{R} \setminus \{0\}$. The parameter c controls the scaling of the identified symplectic structure and Hamiltonian that can be obtained via Legendre transform of L_{imd}. The condition $\ell_{\text{normal}} = 0$ corresponds to fixing the oriented volume of a $2n$-dimensional hypercube $E^{2n} \subset TQ$ to be c. It enforces regularity of the learned Lagrangian. As ℓ_{normal} mostly acts as a non-triviality condition, the weighting θ will typically be small.

Gaussian Processes As an alternative to ANNs, we can fit a Gaussian Process (GP) modelling L_{imd} such that $\text{DEL}_{\text{imd}}\,(q_{j-1}, q_j, q_{j+1}) = 0$ holds on all triples (q_{j-1}, q_j, q_{j+1}) in the training data.

If values $L_{\text{imd}}\,(Z) = (L_{\text{imd}}\,(z_1), \dots, L_{\text{imd}}\,(z_M))$ of a quantity L_{imd} are known at points $Z = (z_1, \dots, z_M) \in TQ^M$ with i.i.d. Gaussian noise at each observation with mean 0 and standard deviation $\sigma \geq 0$, then $L_{\text{imd}}\,(y)$ can be predicted as

$$L_{\text{imd}}\,(y) = k(y, Z)^\top B, \tag{4}$$

where $B = (k(Z, Z) + \sigma I_M)^{-1} L_{\text{imd}}\,(Z)$. Here the prior is a GP with mean 0 and kernel function $k\colon TM \times TM \to \mathbb{R}$, $k(y, Z) = (k(y, z_j))_{j=1}^M$, $k(Z, Z) = (k(z_i, z_j))_{i,j=1}^M$ (see Rasmussen and Williams (2005)).

In case of the midpoint rule, for each pair of consecutive observations of position data (q_a, q_b) in the training data, we form $z = \text{mp}(q_a, q_b) := ((q_b + q_a)/2, (q_b - q_a)/h)$ and collect all such points in the variable $Z \in TQ^M$, where M is the number of pairs. We determine $L_{\text{imd}}\,(Z)$ by imposing that $\text{DEL}_{\text{imd}}\,(q_{j-1}, q_j, q_{j+1}) = 0$ holds on all triples (q_{j-1}, q_j, q_{j+1}) in the training data. For each triple we get the equation

$$\big[\big(\nabla_{\dot{x}} k\,(\text{mp}(q_{j-1}, q_j), Z)^\top - \nabla_{\dot{x}} k\,(\text{mp}(q_j, q_{j+1}), Z)^\top\big)/h + \big(\nabla_x k\,(\text{mp}(q_{j-1}, q_j), Z)^\top - \nabla_x k\,(\text{mp}(q_j, q_{j+1}), Z)^\top\big)/2\big] \cdot B = 0.$$

The normalising condition discussed for ANN is

$$\Big(\frac{1}{2^{2n}} \sum\nolimits_{v \in \text{corners}(E^{2n})} \mathbf{1}_n^\top \nabla_{\dot{q}} k(v, Z)^\top\Big) B = c,$$

where $\mathbf{1}_n$ is a vector of ones. The equations for the triples and the normalising condition yield a linear system of equations for $B = (k(Z, Z) + \sigma I_M)^{-1} L_{\text{imd}}\,(Z)$ whose minimal norm solution is computed. Once B is determined, L_{imd} can be evaluated using (4) and derivatives via automatic differentiation.

Remark 4. Notice that it is not necessary to explicitly perform the inversion $(k(Z, Z) + \sigma I_M)^{-1}$ or to estimate the noise level σ to learn B or to predict L_{imd}. Moreover, assuming correlated noise instead of i.i.d. noise, i.e. replacing the noise matrix σI_M in B with a more complicated matrix, leaves predicted values for L_{imd} invariant.

3. NUMERICAL EXPERIMENTS

In the following experiments, GPs are employed with squared exponential kernel $k(x, y) = \exp(-\|x - y\|^2/\epsilon)$. The parameter ϵ is related to the typical length scale of the problem. As a variational integrator we use the second order accurate midpoint rule.

3.1 Mathematical pendulum

The mathematical pendulum is described by the Lagrangian $L_{\text{ref}} = \frac{1}{2}\dot{q}^2 - \cos(q)$. The energy $H_{\text{ref}} = \frac{1}{2}\dot{q}^2 + \cos(q)$ is conserved along motions. As described in section 2.3, we train the model on position data of 400 trajectories of length 6 and consider the (relatively large) step-size and snapshot time $h = 0.5$.

For comparison, we learn the Lagrangian L directly (rather than L_{imd}): for this, velocity data is approximated using second-order accurate central finite differences. Then a GP is trained using the same technique as in section 2.3, where L_{imd} is replaced by L in the formulas and instances of $\text{mp}(q_j, q_{j+1})$ are replaced by (q_j, v_j), where v_j is the approximated velocity at position q_j. This is referred to as LGP (Lagrangian Gaussian Process) as it constitutes a GP version of Cranmer et al. (2020)'s LNN. For further comparison and an analysis of the different discretisation errors, we also train with exact velocity data for v_j (LGPExact).

Figure 2 shows the first 13 points of a discrete trajectory and its long-term energy behaviour computed with LSI,

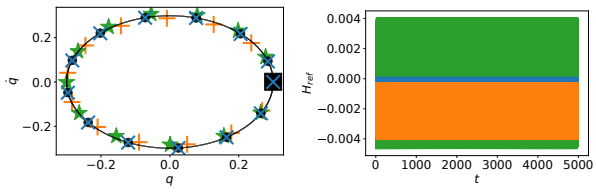


Fig. 2. Left: Pendulum trajectory for LSI (blue, ×), LGP (orange, +), LGPExact (green, ⋆), reference (black, solid). Only the new scheme LSI is accurate since discretisation errors are compensated. Right: Energy along trajectory (colouring as on left hand side).

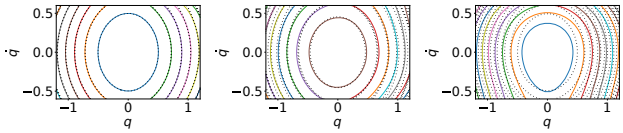


Fig. 3. Left to right: Energy detected by LSI, LGP, and LGPExact. Matching contours of H_{ref} are dotted.

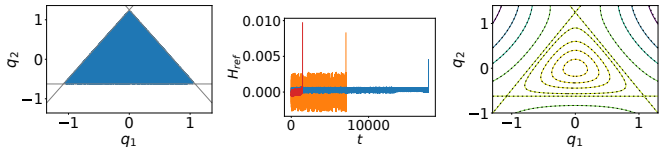


Fig. 4. Left: Position data of an LSI trajectory in a Henon–Heiles system correctly shows a dense orbit. Centre: Energy plots for LSI (blue), LGP (orange), GPFlow (red), Right: Contour plot of the numerical potential and V_{ref} (dashed).

LGP, and LGPExact. As the step-size is quite large, the compensation of discretisation error of the LSI scheme greatly improves accuracy and energy behaviour over methods which try to learn L directly. Even providing exact velocity data does not help as sizeable discretisation errors are introduced when the system is integrated. Interestingly, the energy plot suggests that providing averaged velocity data used to train LGP is better than using exact velocity data in LGPExact: the averaging appears to compensate some of the discretisation errors in the integration step, however not as successfully as our systematic approach LSI.

From the identified L (either computed as $\Lambda(L_{\text{imd}})$ or learned directly) we can calculate the conserved quantity H via Legendre transform of L as $H = q\frac{\partial L}{\partial \dot{q}} - L$. Figure 3 compares the contours of H to the contours of H_{ref}. Clearly, the LSI prediction matches the reference contours best. The computed energy H corresponds to the Hamiltonian of the system which governs the numerical motion. Phase plots of this kind can, therefore, conveniently be used to verify the validity of computations with LSI. Moreover, it shows that LSI can be used for system identification.

3.2 Hénon–Heiles system

The Hénon–Heiles system is governed by $L_{\text{ref}} = \frac{1}{2}\|\dot{q}\|^2 - V_{\text{ref}}(q)$ with potential $V_{\text{ref}}(q) = \frac{1}{2}\|q\|^2 + \mu(q_1^2 q_2 - q_2^3/3)$. We set $\mu = 0.8$, use a step-size and snap-shot time $h = h_{\text{snap}} = 0.1$, and train on 200 trajectories of length 5.

We compare again to LGP with approximated velocities by central finite differences and additionally to a GP representing the flow map on TQ (GPFlow). GPFlow is fitted directly to the position data and the approximated velocity data. GPFlow avoids discretisation errors in the integration step but does not incorporate variational structure. A trajectory and an energy plot of LSI, LGP, and GPFlow is provided in Figure 4. While all trajectories erroneously diverge eventually, LSI's excellent energy preservation properties make the trajectory escape much later. Interestingly, the energy drift of GPFLow causes an early blow-up although at first its absolute energy errors are much smaller than those of LGP which diverges later. This confirms that preservation of variational structure is important. When the potential is computed from $L = \Lambda(L_{\text{imd}})$ its contours match the reference V_{ref} nicely. This verifies the behaviour of numerical solutions of LSI and shows that LSI succeeds in system identification tasks.

4. FUTURE WORK

LSI successfully incorporates variational structure into learning processes of dynamical systems without introducing (biased) discretisation errors. We would like to incorporate symmetries into the learned dynamical system as well, for instance by using symmetric neural network architectures or symmetric kernel functions with Gaussian Processes. In combination with a symmetric variational integrator, the symmetry and variational structure is passed on to the discrete system such that numerical motions conserve the quantities given by Noether's theorem. In this way, completely integrable structure can be preserved under discretisation such that this important structural property is shared by the numerical system.

High order integrators simplify the computation of $L = \Lambda(L_{\text{imd}})$ as correction terms only occur to higher order. However, complicated integrators yield more complicated loss functions ℓ_{data} (ANN) or more involved linear systems (GP based method). We would like to investigate these trade-offs and the potential of using different variational integrators within LSI. Another direction of research is to incorporate external forces into the considered systems.

Moreover, we would like to employ the statistical framework of Gaussian Processes to investigate how uncertainty in the training data relates to model uncertainty in the learned Lagrangian and in the computed trajectories.

REFERENCES

Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. (2020). Lagrangian Neural Networks. URL `https://arxiv.org/abs/2003.04630`.

Ober-Blöbaum, S. and Offen, C. (2022). Variational integration of learned dynamical systems. URL `https://arxiv.org/abs/2112.12619`.

Rasmussen, C.E. and Williams, C.K.I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press.

Takehiro Aoshima, Takashi Matsubara, T.Y. (2021). Deep discrete-time Lagrangian mechanics. *ICLR 2021 SimDL Workshop*.

Vermeeren, M. (2017). Modified equations for variational integrators. *Numerische Mathematik*, 137(4), 1001–1037. doi:10.1007/s00211-017-0896-4.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Maximum Entropy Optimal Density Control of Discrete-time Linear Systems

Kaito Ito * **Kenji Kashima** *

* *G*
k *k* *; kk* *k*

Abstract: Entropy regularization, or a maximum entropy method for optimal control has attracted much attention especially in reinforcement learning due to its many advantages such as a natural exploration strategy and robustness against disturbances. Nevertheless, for safety-critical applications, it is crucial to suppress state uncertainty due to the stochasticity of high-entropy control policies and dynamics to an acceptable level. To achieve this, we consider the problem of steering a state distribution of a deterministic discrete-time linear system to a specified one at final time with entropy-regularized minimum energy control. We show that this problem boils down to solving coupled Lyapunov equations. Based on this, we derive the existence, uniqueness, and explicit form of the optimal policy.

Optimal control, stochastic control, maximum entropy, discrete-time linear system

1. INTRODUCTION

Optimal control theory is a powerful mathematical tool for achieving control objectives while considering, for example, energy efficiency (Lewis et al., 2012). Recently, there has been considerable interest in maximum entropy optimal control (MaxEnt OC) especially in reinforcement learning (RL) (Haarnoja et al., 2017, 2018; Levine, 2018; Ho and Ermon, 2016). MaxEnt OC seeks to optimize an objective function which includes an additional entropy regularization term for control policies. This framework offers many advantages such as performing good exploration for RL (Haarnoja et al., 2017), robustness against disturbances (Eysenbach and Levine, 2021), and equivalence between MaxEnt OC and an inference problem (Levine, 2018), to name a few.

On the other hand, in many safety-critical applications, it is important to limit state uncertainty due to the stochasticity of high-entropy policies and dynamics to an acceptable level. A straightforward approach to achieve this is to impose a hard constraint in the state distribution at a specified time. Steering the state of a dynamical system to a desired distribution without entropy regularization has been addressed in the literature. In (Chen et al., 2016), the problem of steering a Gaussian initial density of a continuous-time linear system to a final one with minimum energy is considered, and the optimal policy is derived in explicit form. In (Goldshtein and Tsiotras, 2017), the above problem for a discrete-time linear system is investigated, and the optimality condition for a linear controller gain is derived. See also (Chen et al., 2021) for an extensive review of this area.

In this paper, we tackle the entropy-regularized minimum energy density control problem for deterministic discrete-time linear systems. We reveal that this problem boils down to solving two Lyapunov difference equations coupled through their boundary values. Our main contribution is to show the existence and uniqueness of the optimal policy and then derive its explicit form.

This paper is organized as follows: In Section 2, we provide the problem formulation and derive the coupled Lyapunov equations. The existence, uniqueness, and explicit form of the optimal policy are given in Section 3. Some concluding remarks are given in Section 4.

Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{Z}_{>0}$ denote the set of positive integers. The set of integers $\{k, k+1, \dots, l\}$ $(k < l)$ is denoted by $[\![k, l]\!]$. Denote by $\mathcal{S}^n$ the set of all real symmetric $n \times n$ matrices. For a matrix $A \in \mathcal{S}^n$, we write $A \succ 0$ (resp. $A \prec 0$) if A is positive (resp. negative) definite. For $A \succ 0$, $A^{1/2}$ denotes the unique positive definite square root. The identity matrix is denoted by I, and its dimension depends on the context. The Euclidean norm is denoted by $\|\cdot\|$. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a complete probability space and $\mathbb{E}$ be the expectation with respect to $\mathbb{P}$. For an $\mathbb{R}^n$-valued random vector w, $w \sim \mathcal{N}(\mu, \Sigma)$ means that w has a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^n$ and covariance matrix Σ. When $\Sigma \succ 0$, the density function of $w \sim \mathcal{N}(\mu, \Sigma)$ is denoted by $\mathcal{N}(\cdot|\mu, \Sigma)$.

2. PROBLEM FORMULATION AND PRELIMINARY ANALYSIS

In this paper, we consider the following optimal control problem.

Given a finite horizon $N \in \mathbb{Z}_{>0}$, find a policy $\pi = \{\pi_k\}_{k=0}^{N-1}$ that solves

$$\underset{\pi}{\text{minimize}} \quad \mathbb{E}\left[\sum_{k=0}^{N-1}\left(\frac{1}{2}\|u_k\|^2 - \varepsilon \mathrm{H}(\pi_k(\cdot|x_k))\right)\right] \tag{1}$$

$$\text{subject to} \quad x_{k+1} = A_k x_k + B_k u_k, \tag{2}$$

$$u_k \sim \pi_k(\cdot|x_k), \tag{3}$$

$$x_0 \sim \mathcal{N}(0, \bar{\Sigma}_0),\ x_N \sim \mathcal{N}(0, \bar{\Sigma}_N), \tag{4}$$

where $\varepsilon > 0,\ \bar{\Sigma}_0, \bar{\Sigma}_N \succ 0,\ x_k \in \mathbb{R}^n, u_k \in \mathbb{R}^m$. A stochastic policy $\pi_k(\cdot|x)$ denotes the conditional density of u_k given $x_k = x$, and $\mathrm{H}(\pi_k(\cdot|x)) := -\int_{\mathbb{R}^m} \pi_k(u|x) \log \pi_k(u|x) \mathrm{d}u$ denotes the entropy of $\pi_k(\cdot|x)$. ♢

To tackle the above problem, let us start by introducing the auxiliary problem.

Given a finite horizon $N \in \mathbb{Z}_{>0}$, find a policy $\pi = \{\pi_k\}_{k=0}^{N-1}$ that solves

$$\begin{aligned} \underset{\pi}{\text{minimize}} \quad & \mathbb{E}\left[\frac{1}{2} x_N^\top \Pi_N x_N \right. \\ & \left. + \sum_{k=0}^{N-1} \left(\frac{1}{2}\|u_k\|^2 - \varepsilon \mathrm{H}(\pi_k(\cdot|x_k))\right)\right] \quad (5) \\ \text{subject to} \quad & x_{k+1} = A_k x_k + B_k u_k,\ u_k \sim \pi_k(\cdot|x_k), \\ & x_0 \sim \mathcal{N}(0, \bar{\Sigma}_0), \end{aligned}$$

where $\Pi_N \in \mathcal{S}^n$. ♢

Problem 2 has a terminal cost instead of a constraint on the density of the final state. Similarly to the conventional LQR problem (Lewis et al., 2012), we can obtain the optimal policy for Problem 2 explicitly. Due to the limited space, we omit the proof.

Assume that $\Pi_k \in \mathcal{S}^n$ satisfies $I + B_k^\top \Pi_{k+1} B_k \succ 0$ for any $k \in [\![0, N-1]\!]$ and is a solution of the following Riccati difference equation:

$$\begin{aligned} \Pi_k = A_k^\top \Pi_{k+1} A_k - A_k^\top \Pi_{k+1} B_k (I + B_k^\top \Pi_{k+1} B_k)^{-1} \\ \times B_k^\top \Pi_{k+1} A_k,\ k \in [\![0, N-1]\!]. \quad (6) \end{aligned}$$

Then, the unique optimal policy for Problem 2 is given by

$$\begin{aligned} \pi_k^*(u|x) = \mathcal{N}\big(u \big| -(I + B_k^\top \Pi_{k+1} B_k)^{-1} B_k^\top \Pi_{k+1} A_k x, \\ \varepsilon (I + B_k^\top \Pi_{k+1} B_k)^{-1}\big), \\ x \in \mathbb{R}^n, u \in \mathbb{R}^m, k \in [\![0, N-1]\!]. \quad (7) \end{aligned}$$

♢

The system (2) driven by the policy (7) is given by

$$x_{k+1}^* = \bar{A}_k x_k^* + B_k w_k^*,\ w_k^* \sim \mathcal{N}(0, \varepsilon (I + B_k^\top \Pi_{k+1} B_k)^{-1}), \quad (8)$$

$$\bar{A}_k := A_k - B_k (I + B_k^\top \Pi_{k+1} B_k)^{-1} B_k^\top \Pi_{k+1} A_k \quad (9)$$

where $\{w_k^*\}$ is an independent sequence. Suppose that $\Sigma_k := \mathbb{E}\left[x_k^* (x_k^*)^\top\right]$ satisfies $\Sigma_0 = \bar{\Sigma}_0, \Sigma_N = \bar{\Sigma}_N$. Then, from Proposition 1, the policy (7) is the unique optimal solution of Problem 1. This is because for any policy satisfying (4), the terminal cost $\mathbb{E}[x_N^\top \Pi_N x_N/2]$ takes the same value. From (8), Σ_k evolves as

$$\Sigma_{k+1} = \bar{A}_k \Sigma_k \bar{A}_k^\top + \varepsilon B_k (I + B_k^\top \Pi_{k+1} B_k)^{-1} B_k^\top. \quad (10)$$

Note that

$$\begin{aligned} \bar{A}_k \Sigma_k \bar{A}_k^\top = (I + B_k B_k^\top \Pi_{k+1})^{-1} A_k \Sigma_k A_k^\top \\ \times (I + B_k B_k^\top \Pi_{k+1})^{-\top}. \quad (11) \end{aligned}$$

Therefore, if $\Sigma_0 \succ 0$ and A_k is invertible for any $k \in [\![0, N-1]\!]$, it holds $\Sigma_k \succ 0$ for any $k \in [\![1, N]\!]$. Henceforth, we assume the invertibility of A_k.

Now, inspired by (Chen et al. (2016)), we introduce $H_k := \varepsilon \Sigma_k^{-1} - \Pi_k$. Assume that Π_k and H_k are invertible on the time interval $[\![0, N]\!]$. Noting that

$$(\Pi_k + H_k)^{-1} = \Pi_k^{-1} - \Pi_k^{-1} (H_k^{-1} + \Pi_k^{-1})^{-1} \Pi_k^{-1},$$

we obtain

$$\Pi_k^{-1} + H_k^{-1} = \left(\Pi_k - \frac{1}{\varepsilon} \Pi_k \Sigma_k \Pi_k\right)^{-1}. \quad (12)$$

In addition, a straightforward calculation with (6) and (10) shows that

$$\begin{aligned} & A_k^{-1} (\Pi_{k+1}^{-1} + H_{k+1}^{-1}) A_k^{-\top} \\ & = \left(A_k^\top \left(\Pi_{k+1} - \frac{1}{\varepsilon} \Pi_{k+1} \Sigma_{k+1} \Pi_{k+1}\right) A_k\right)^{-1} \\ & = \left(\Pi_k - \frac{1}{\varepsilon} \Pi_k \Sigma_k \Pi_k\right)^{-1} = \Pi_k^{-1} + H_k^{-1}. \quad (13) \end{aligned}$$

Hence, we have

$$\Pi_{k+1}^{-1} + H_{k+1}^{-1} = A_k (\Pi_k^{-1} + H_k^{-1}) A_k^\top. \quad (14)$$

Moreover, the Riccati equation (6) can be rewritten as

$$\Pi_{k+1}^{-1} = A_k \Pi_k^{-1} A_k^\top - B_k B_k^\top. \quad (15)$$

From (14) and (15), it holds

$$H_{k+1}^{-1} = A_k H_k^{-1} A_k^\top + B_k B_k^\top. \quad (16)$$

Therefore, $P_k := H_k^{-1}$ and $Q_k := \Pi_k^{-1}$ satisfy the Lyapunov difference equations

$$P_{k+1} = A_k P_k A_k^\top + B_k B_k^\top, \quad (17)$$

$$Q_{k+1} = A_k Q_k A_k^\top - B_k B_k^\top \quad (18)$$

for $k \in [\![0, N-1]\!]$, and the boundary conditions $\Sigma_0 = \bar{\Sigma}_0, \Sigma_N = \bar{\Sigma}_N$ are written as

$$\varepsilon \bar{\Sigma}_0^{-1} = P_0^{-1} + Q_0^{-1}, \quad (19)$$

$$\varepsilon \bar{\Sigma}_N^{-1} = P_N^{-1} + Q_N^{-1}. \quad (20)$$

In summary, we obtain the following proposition.

Assume that for any $k \in [\![0, N-1]\!]$, A_k is invertible. Assume further that P_k and Q_k satisfy the equations (17), (18) with the boundary conditions (19),(20) and are invertible on $[\![0, N]\!]$, and that it holds $I + B_k^\top Q_{k+1}^{-1} B_k \succ 0$ for any $k \in [\![0, N-1]\!]$. Then, the policy (7) with $\Pi_k = Q_k^{-1}$ is the unique optimal policy for Problem 1. ♢

3. SOLUTION OF MAXIMUM ENTROPY OPTIMAL DENSITY CONTROL PROBLEM

In this section, we analyze the Lyapunov equations (17),(18) with the boundary conditions (19),(20). Under the invertibility of A_k, define the state-transition matrix

$$\Phi(k, l) := \begin{cases} A_{k-1} A_{k-2} \cdots A_l, & k > l \geq 0 \\ I, & k = l \geq 0 \\ A_k^{-1} A_{k+1}^{-1} \cdots A_{l-1}^{-1}, & 0 \leq k < l \end{cases}, \quad (21)$$

and introduce the reachability Gramian

$$G_{\mathrm{r}}(k_1, k_0) := \sum_{k=k_0}^{k_1-1} \Phi(k_1, k+1) B_k B_k^\top \Phi(k_1, k+1)^\top,\ k_0 < k_1, \quad (22)$$

and the controllability Gramian

$$G_{\mathrm{c}}(k_1, k_0) := \sum_{k=k_0}^{k_1-1} \Phi(k_0, k+1) B_k B_k^\top \Phi(k_0, k+1)^\top,\ k_0 < k_1. \quad (23)$$

Note that since $G_{\mathrm{c}}(k_1, k_0) = \Phi(k_0, k_1) G_{\mathrm{r}}(k_1, k_0) \Phi(k_0, k_1)^\top$, if $G_{\mathrm{r}}(k_1, k_0)$ is invertible, $G_{\mathrm{c}}(k_1, k_0)$ is also invertible. Now, we provide the solutions of (17),(18) with (19),(20).

Assume that for any $k \in [\![0, N-1]\!]$, A_k is invertible, and there exists $k_{\mathrm{r}} \in [\![1, N-1]\!]$ such that $G_{\mathrm{r}}(k,0)$ is invertible for any $k \in [\![k_{\mathrm{r}}, N]\!]$ and $G_{\mathrm{r}}(N,k)$ is invertible for any $k \in [\![0, k_{\mathrm{r}}-1]\!]$. Assume further that for

$$S_0 := \frac{1}{\varepsilon} G_{\mathrm{c}}(N,0)^{-\frac{1}{2}} \bar{\Sigma}_0 G_{\mathrm{c}}(N,0)^{-\frac{1}{2}}, \tag{24}$$

$$S_N := \frac{1}{\varepsilon} G_{\mathrm{c}}(N,0)^{-\frac{1}{2}} \Phi(0,N) \bar{\Sigma}_N \Phi(0,N)^\top G_{\mathrm{c}}(N,0)^{-\frac{1}{2}}, \tag{25}$$

the following two matrices are invertible.

$$\mathcal{F}(S_0, S_N) := S_0 + \frac{1}{2} I - \left(S_0^{\frac{1}{2}} S_N S_0^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}}, \tag{26}$$

$$\mathcal{B}(S_0, S_N) := -S_0 + \frac{1}{2} I + \left(S_0^{\frac{1}{2}} S_N S_0^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}}. \tag{27}$$

Then, the equations (17),(18) with the boundary conditions (19),(20) have two sets of solutions $(P_{\pm,k}, Q_{\pm,k})$, $k \in [\![0, N]\!]$ specified by

$$Q_{\pm,0} = G_{\mathrm{c}}(N,0)^{\frac{1}{2}} S_0^{\frac{1}{2}} \left(S_0 + \frac{1}{2} I \pm \left(S_0^{\frac{1}{2}} S_N S_0^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} \right)^{-1} \times S_0^{\frac{1}{2}} G_{\mathrm{c}}(N,0)^{\frac{1}{2}}, \tag{28}$$

$$P_{\pm,0} = (\varepsilon \bar{\Sigma}_0^{-1} - Q_{\pm,0}^{-1})^{-1}. \tag{29}$$

In addition, the two sets of solutions $(P_{\pm,k}, Q_{\pm,k})$ have the following properties.

(i) $P_{-,k}$ and $Q_{-,k}$ are both invertible on $[\![0, N]\!]$, and for any $k \in [\![0, N-1]\!]$, it holds $I + B_k^\top Q_{-,k+1}^{-1} B_k \succ 0$;

(ii) If $Q_{+,k}$ is invertible on $[\![0, N]\!]$, there exists $s \in [\![0, N-1]\!]$ such that $I + B_s^\top Q_{+,s+1}^{-1} B_s$ is not positive definite.

Proof. First, introduce the change of variables $\xi_k := G_{\mathrm{c}}(N,0)^{-1/2} \Phi(0,k) x_k$. Then the system (2) is transformed into

$$\xi_{k+1} = \xi_k + \underbrace{G_{\mathrm{c}}(N,0)^{-1/2} \Phi(0,k+1) B_k}_{=: B_{\mathrm{new},k}} u_k. \tag{30}$$

We will prove the statement in this new set of coordinates and then turn back to the original set of coordinates at the end. The Lyapunov equations associated with the transformed system (30) are given by

$$P_{\mathrm{new},k+1} = P_{\mathrm{new},k} + B_{\mathrm{new},k} B_{\mathrm{new},k}^\top, \tag{31}$$

$$Q_{\mathrm{new},k+1} = Q_{\mathrm{new},k} - B_{\mathrm{new},k} B_{\mathrm{new},k}^\top. \tag{32}$$

The relationship between $Q_{\mathrm{new},k}$ and Q_k is as follows.

$$Q_{\mathrm{new},k} = G_{\mathrm{c}}(N,0)^{-\frac{1}{2}} \Phi(0,k) Q_k \Phi(0,k)^\top G_{\mathrm{c}}(N,0)^{-\frac{1}{2}}. \tag{33}$$

Indeed, substituting (33) into $Q_{\mathrm{new},k+1} - Q_{\mathrm{new},k}$ yields

$$\begin{aligned} Q_{\mathrm{new},k+1} - Q_{\mathrm{new},k} &= G_{\mathrm{c}}(N,0)^{-\frac{1}{2}} \Phi(0,k) \\ &\quad \times (A_k^{-1} Q_{k+1} A_k^{-\top} - Q_k) \Phi(0,k)^\top G_{\mathrm{c}}(N,0)^{-\frac{1}{2}} \\ &= -G_{\mathrm{c}}(N,0)^{-\frac{1}{2}} \Phi(0,k+1) B_k B_k^\top \Phi(0,k+1)^\top G_{\mathrm{c}}(N,0)^{-\frac{1}{2}} \\ &= -B_{\mathrm{new},k} B_{\mathrm{new},k}^\top, \end{aligned} \tag{34}$$

which coincides with (32). The controllability Gramian and the reachability Gramian corresponding to ξ_k are given by

$$G_{\mathrm{c,new}}(k_1, k_0) = G_{\mathrm{r,new}}(k_1, k_0) := \sum_{k=k_0}^{k_1-1} B_{\mathrm{new},k} B_{\mathrm{new},k}^\top \tag{35}$$

satisfying $G_{\mathrm{c,new}}(N,0) = G_{\mathrm{r,new}}(N,0) = I$. The initial and final covariance matrices for ξ_0 and ξ_N are given by

$$\bar{\Sigma}_{\mathrm{new},0} := G_{\mathrm{c}}(N,0)^{-1/2} \bar{\Sigma}_0 G_{\mathrm{c}}(N,0)^{-1/2} = \varepsilon S_0, \tag{36}$$

$$\begin{aligned} \bar{\Sigma}_{\mathrm{new},N} &:= G_{\mathrm{c}}(N,0)^{-1/2} \Phi(0,N) \bar{\Sigma}_N \Phi(0,N)^\top G_{\mathrm{c}}(N,0)^{-1/2} \\ &= \varepsilon S_N. \end{aligned} \tag{37}$$

In what follows, to simplify notation, we omit the subscript "new." First, by substituting

$$P_N = P_0 + I, \ Q_N = Q_0 - I$$

into the boundary conditions (19),(20), similarly to the proof of (Chen et al., 2016, Proposition 4), we obtain the two sets of initial values

$$\begin{aligned} Q_{\pm,0} &= S_0^{\frac{1}{2}} \left(S_0 + \frac{1}{2} I \pm \left(S_0^{\frac{1}{2}} S_N S_0^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} \right)^{-1} S_0^{\frac{1}{2}}, \\ P_{\pm,0} &= (S_0^{-1} - Q_{\pm,0}^{-1})^{-1} \end{aligned}$$

under the invertibility of $\mathcal{F}(S_0, S_N)$ and $\mathcal{B}(S_0, S_N)$. Next, we show that $Q_{-,k} = Q_{-,0} - G_{\mathrm{r}}(k,0)$ is invertible on $[\![0, N]\!]$. Note that formally

$$\begin{aligned} &(Q_{-,0} - G_{\mathrm{r}}(k,0))^{-1} = -G_{\mathrm{r}}(k,0)^{-1} \\ &\quad - G_{\mathrm{r}}(k,0)^{-1} (Q_{-,0}^{-1} - G_{\mathrm{r}}(k,0)^{-1})^{-1} G_{\mathrm{r}}(k,0)^{-1} \\ &= -G_{\mathrm{r}}(k,0)^{-1} - G_{\mathrm{r}}(k,0)^{-1} S_0^{\frac{1}{2}} \Bigg[S_0 + \frac{1}{2} I \\ &\quad - \left(S_0^{\frac{1}{2}} S_N S_0^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} - S_0^{\frac{1}{2}} G_{\mathrm{r}}(k,0)^{-1} S_0^{\frac{1}{2}} \Bigg]^{-1} S_0^{\frac{1}{2}} G_{\mathrm{r}}(k,0)^{-1}. \end{aligned} \tag{38}$$

By assumption, for any $k \in [\![k_{\mathrm{r}}, N]\!]$, $G_{\mathrm{r}}(k,0)$ is invertible. The term in the square brackets obviously attains its maximum at $k = N$

$$\frac{1}{2} I - \left(S_0^{1/2} S_N S_0^{1/2} + \frac{1}{4} I \right)^{1/2} \prec 0.$$

Therefore the term in the brackets is invertible for any $k \in [\![k_{\mathrm{r}}, N]\!]$. This implies that $Q_{-,k}$ has the inverse matrix (38). On the other hand, $Q_{-,k}$ also admits the expression $Q_{-,k} = Q_{-,N} + G_{\mathrm{r}}(N,k)$. Hence, the inverse matrix is formally

$$\begin{aligned} &(Q_{-,N} + G_{\mathrm{r}}(N,k))^{-1} = G_{\mathrm{r}}(N,k)^{-1} \\ &\quad - G_{\mathrm{r}}(N,k)^{-1} (Q_{-,N}^{-1} + G_{\mathrm{r}}(N,k)^{-1})^{-1} G_{\mathrm{r}}(N,k)^{-1} \\ &= G_{\mathrm{r}}(N,k)^{-1} - G_{\mathrm{r}}(N,k)^{-1} ((Q_{-,0} - I)^{-1} + G_{\mathrm{r}}(N,k)^{-1})^{-1} \\ &\quad \times G_{\mathrm{r}}(N,k)^{-1} \\ &= G_{\mathrm{r}}(N,k)^{-1} - G_{\mathrm{r}}(N,k)^{-1} S_0^{\frac{1}{2}} \Bigg[-S_0 + \Bigg(-\frac{1}{2} I + \Bigg(S_0^{\frac{1}{2}} S_N S_0^{\frac{1}{2}} \\ &\quad + \frac{1}{4} I \Bigg)^{\frac{1}{2}} \Bigg)^{-1} + S_0^{\frac{1}{2}} G_{\mathrm{r}}(N,k)^{-1} S_0^{\frac{1}{2}} \Bigg]^{-1} S_0^{\frac{1}{2}} G_{\mathrm{r}}(N,k)^{-1}. \end{aligned} \tag{39}$$

Then by the same argument as for the time interval $[\![k_{\mathrm{r}}, N]\!]$, $Q_{-,k}$ is also invertible on $[\![0, k_{\mathrm{r}}-1]\!]$. Similarly, it can be shown that $P_{-,k}$ is invertible on $[\![0, N]\!]$.

Next, we prove that $I + B_k^\top Q_{-,k+1}^{-1} B_k \succ 0$ for any $k \in [\![0, N-1]\!]$. Note that

$$\begin{aligned} (I + B_k^\top Q_{-,k+1}^{-1} B_k)^{-1} &= I - B_k^\top (Q_{-,k+1} + B_k B_k^\top)^{-1} B_k \\ &= I - B_k^\top Q_{-,k}^{-1} B_k. \end{aligned}$$

Hence, we show $I - B_k^\top Q_{-,k}^{-1} B_k \succ 0$. From (38), for $k \in [\![k_\mathrm{r}, N-1]\!]$, we have

$$\begin{aligned} I - B_k^\top Q_{-,k}^{-1} B_k &= I + B_k^\top G_\mathrm{r}(k,0)^{-1} B_k \\ &\quad + B_k^\top G_\mathrm{r}(k,0)^{-1} S_0^{\frac{1}{2}} \Bigg[S_0 + \frac{1}{2} I - \left(S_0^{\frac{1}{2}} S_N S_0^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} \\ &\quad - S_0^{\frac{1}{2}} G_\mathrm{r}(k,0)^{-1} S_0^{\frac{1}{2}} \Bigg]^{-1} S_0^{\frac{1}{2}} G_\mathrm{r}(k,0)^{-1} B_k. \end{aligned} \tag{40}$$

Since the expression in the square brackets is negative definite on $[\![k_\mathrm{r}, N-1]\!]$, it holds for sufficiently small $\delta \in (0,1)$,

$$\begin{aligned} &I - B_k^\top Q_{-,k}^{-1} B_k \succ I + B_k^\top G_\mathrm{r}(k,0)^{-1} B_k + B_k^\top G_\mathrm{r}(k,0)^{-1} S_0^{\frac{1}{2}} \\ &\times \left[S_0 - \delta S_0 - S_0^{\frac{1}{2}} G_\mathrm{r}(k,0)^{-1} S_0^{\frac{1}{2}} \right]^{-1} S_0^{\frac{1}{2}} G_\mathrm{r}(k,0)^{-1} B_k. \end{aligned}$$

Hence, we get

$$\begin{aligned} &I - B_k^\top Q_{-,k}^{-1} B_k \succ I + B_k^\top (G_\mathrm{r}(k,0) - (1-\delta)^{-1} I)^{-1} B_k \\ &= I - B_k^\top \left(\sum_{s=k}^{N-1} B_s B_s^\top + \left((1-\delta)^{-1} - 1 \right) I \right)^{-1} B_k. \end{aligned} \tag{41}$$

In addition, it holds that

$$\begin{aligned} &\left(I - B_k^\top \left(\sum_{s=k}^{N-1} B_s B_s^\top + \left((1-\delta)^{-1} - 1 \right) I \right)^{-1} B_k \right)^{-1} \\ &= I + B_k^\top \left(\sum_{s=k+1}^{N-1} B_s B_s^\top + \left((1-\delta)^{-1} - 1 \right) I \right)^{-1} B_k \succ 0. \end{aligned}$$

Consequently, we obtain $I + B_k^\top Q_{-,k+1}^{-1} B_k \succ 0$ for $k \in [\![k_\mathrm{r}, N-1]\!]$. Noting also that $Q_{-,k}^{-1}$ is given by (39) for $k \in [\![0, k_\mathrm{r}-1]\!]$, by the same argument above, it can be shown that $I + B_k^\top Q_{-,k+1}^{-1} B_k \succ 0$ for $k \in [\![0, k_\mathrm{r}-1]\!]$.

Next, we show the property (ii). Note that since $0 \prec Q_{+,0} \prec I$, it holds $Q_{+,N} \prec 0$, and thus there exists $s \in [\![0, N-1]\!]$ such that $Q_{+,s} \succ 0$ and $Q_{+,s+1}$ is not positive definite. By assumption, $Q_{+,s+1}$ is invertible and

$$\begin{aligned} Q_{+,s+1}^{-1} &= (Q_{+,s} - B_s B_s^\top)^{-1} \\ &= Q_{+,s}^{-1} + Q_{+,s}^{-1} B_s (I - B_s^\top Q_{+,s}^{-1} B_s)^{-1} B_s^\top Q_{+,s}^{-1}. \end{aligned} \tag{42}$$

Now assume $I - B_s^\top Q_{+,s}^{-1} B_s \succ 0$. Then by $Q_{+,s}^{-1} \succ 0$ and (42), we have $Q_{+,s+1}^{-1} \succ 0$, which contradicts the fact that $Q_{+,s+1}$ is not positive definite. Combining this with $I + B_s^\top Q_{+,s+1}^{-1} B_s = (I - B_s^\top Q_{+,s}^{-1} B_s)^{-1}$, we obtain (ii).

Finally, by employing the relationship (33), we obtain the desired result. □

Note that the corresponding result (Chen et al., 2016, Proposition 4) for the continuous-time system also requires the invertibility of $\mathcal{F}(S_0, S_N)$ and $\mathcal{B}(S_0, S_N)$, which is not mentioned in the statement.

By Propositions 2 and 3, we come to the main result of this paper.

Suppose that the assumptions of Proposition 3 are satisfied. Then, the unique optimal policy for Problem 1 is given by

$$\begin{aligned} \pi_k(u|x) = \mathcal{N}\Big(&u \,|\, -(I + B_k^\top Q_{-,k+1}^{-1} B_k)^{-1} B_k^\top Q_{-,k+1}^{-1} A_k x, \\ &\varepsilon (I + B_k^\top Q_{-,k+1}^{-1} B_k)^{-1} \Big), \\ &x \in \mathbb{R}^n, u \in \mathbb{R}^m, k \in [\![0, N-1]\!], \end{aligned} \tag{43}$$

where $Q_{-,k}$ is a solution of (18) with the initial value $Q_{-,0}$ in (28). ♢

4. CONCLUSION

In this paper, we analyzed maximum entropy optimal density control of deterministic discrete-time linear systems. In particular, we revealed the existence and uniqueness of the optimal solution, and provided the explicit construction of the optimal policy. Future work includes removing or relaxing the invertibility assumption on $A_k, \mathcal{F}(S_0, S_N), \mathcal{B}(S_0, S_N)$ and exploring the connection between our result and Schrödinger bridges (Beghi, 1996). Another direction of future work is to extend the result to the case where a quadratic state cost is also present.

ACKNOWLEDGEMENTS

This work was supported in part by JSPS KAKENHI under Grant Number JP21J14577, JP21H04875, and by JST, ACT-X under Grant Number JPMJAX2102.

REFERENCES

Beghi, A. (1996). On the relative entropy of discrete-time Markov processes with given end-point densities. , 42(5), 1529–1535.

Chen, Y., Georgiou, T.T., and Pavon, M. (2016). Optimal steering of a linear stochastic system to a final probability distribution, Part I. , 61(5), 1158–1169.

Chen, Y., Georgiou, T.T., and Pavon, M. (2021). Optimal transport in systems and control. , 4, 89–113.

Eysenbach, B. and Levine, S. (2021). Maximum entropy RL (provably) solves some robust RL problems. .

Goldshtein, M. and Tsiotras, P. (2017). Finite-horizon covariance control of linear time-varying systems. In , 3606–3611. IEEE.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In , 1352–1361. PMLR.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In , 1861–1870. PMLR.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. , 29.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. .

Lewis, F.L., Vrabie, D., and Syrmos, V.L. (2012). . John Wiley & Sons.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

INTERPOLATING MATRICES

ALBERTO DAYAN - EXTENDED ABSTRACT

This talk is based on the works in [5], [6] and [7], and it aims to extend some well known results on interpolating sequences to sequences of matrices of arbitrary dimensions. First, we will consider the case of a sequence of square matrices with spectra in the unit disc, and we will obtain an analogue of Carleson's interpolation Theorems, [3] and [4], to this setting. The second part of the talk will consider the case of interpolating d-tuples of matrices: the main difference with the scalar setting is that the components of a d-tuple of matrices need not, in general, to commute. We will see how an analogue of the Pick property enjoyed by the NC (non commutative) Drury-Arveson space allows one to characterize interpolating sequences on the NC unit ball in terms of some Riesz system-type conditions on NC kernels. We will discuss examples, and some possible directions for some future research in the topic.

The one variable case

The first step is to define what an interpolating sequence of matrices is. Since we evaluate an H^∞ function at a matrix via the Riesz-Dunford functional calculus, the first requirement for a sequences $A = (A_n)_{n\in\mathbb{N}}$ of matrices to be interpolating is that their spectra lie in the unit disc. As for a concrete definition for A to be interpolating, a rather versatile choice is the following:

Definition 1 (Interpolating Matrices)**.** *A sequence of matrices $(A_n)_{n\in\mathbb{N}}$ with spectra in the unit disc is interpolating if, for any bounded sequence $(\phi_n)_{n\in\mathbb{N}}$ in H^∞, there exists a function ϕ in H^∞ so that $\phi(A_n) = \phi_n(A_n)$, for any n in $\mathbb{N}$.*

If we chose $(\phi_n)_{n\in\mathbb{N}}$ to be made of constant functions it is immediate to see that this extends the classic definition of interpolating sequences of scalars.
In order to characterize interpolating sequences of matrices, one has to be able to *separate* them. This can be done by using the Hardy space H^2, the reproducing kernel Hilbert space of holomorphic functions on the unit disc with square summable Taylor coefficients. Given a sequence of matrices A, one can define, for any n in $\mathbb{N}$,

$$H_n := \{f \in \mathrm{H}^2 \mid f(A_n) = 0\}^\perp.$$

Each H_n is, in fact, a **model space** in H^2. In particular, it is the orthogonal complement in H^2 of the set of all multiples of a finite Blaschke product B_n. If A was a sequence of scalars $(\lambda_n)_{n\in\mathbb{N}}$, each H_n would be a one dimensional subspace of H^2 spanned by s_λ, the normalized Szegö kernel at λ_n. Looking for a way to separate the sequence $H = (H_n)_{n\in\mathbb{N}}$, we say that H is

- **strongly separated** if the sine of the angle between an element of H and the closure of the span of all the others subspaces in H is uniformly bounded below;
- **weakly separated** if the sine of the angle between any two distinct elements of H is uniformly bounded below;

Date: July 14, 2022.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

- a **Riesz system** if there exists a $C \geq 1$ so that, however we choose a sequence of unit vectors $(x_n)_{n\in\mathbb{N}}$ so that x_n belongs to H_n for any n in $\mathbb{N}$, then for any $a = (a_n)_{n\in\mathbb{N}}$ in l^2

$$\frac{1}{C}||a||_{l^2} \leq \left|\left|\sum_{n\in\mathbb{N}} a_n x_n\right|\right|_{\mathrm{H}^2} \leq C||a||_{l^2}.$$

  The least constant C for which this holds is called the *Riez bound*. If the right inequality holds, then H is a **Bessel system**.

In [1, Ch. 9] one can find a re-statement of the celebrated characterization of interpolating sequences due to Carleson in a operator theoretical language: in [3] he proved that a sequence of scalars $\Lambda = (\lambda_n)_{n\in\mathbb{N}}$ is interpolating if and only if $(s_{\lambda_n})_{n\in\mathbb{N}}$ is strongly separated. Later on, [4], he proved that Λ is interpolating if and only if $(s_{\lambda_n})_{n\in\mathbb{N}}$ is a weakly separated Bessel system in H^2. Moreover, Shapiro and Shields showed in [9] that Λ is interpolating if and only if $(s_{\lambda_n})_{n\in\mathbb{N}}$ is a Riesz system. The first main result of this talk reads as follow:

Theorem 1. *Let A be a sequence of matrices with spectra in the unit disc, and let H be the associated sequence of model spaces in H^2. Then the following are equivalent:*

(i): *A is interpolating;*
(ii): *H is strongly separated;*
(iii): *H is a Riesz system in H^2.*
(iv): *H is a weakly separated Bessel system.*

Interpolating d-tuples

Interpolating sequences are significantly less understood in the multi-variable setting. Looking for extending some of the few known results on the topic to sequences of *d-tuples* of matrices, a first crucial observation is that the components of such d-tuples might not commute, unless they are one-dimensional. Therefore, one has to work with the well developed theory of *noncommutative analytic functions*, and in particular with the nc (non commutative) reproducing kernel Hilbert spaces defined in [2]. We can think of an nc analytic function as a *noncommuative power series*

$$f(z) = \sum_{k\in\mathbb{W}_d} a_k z^k,$$

where $z = (z_1, \ldots, z_d)$ is a d-tuple of non commuting variables, and $\mathbb{W}_d$ is the set of all free words with d generators. In [8] Salomon, Shalit and Shamovich considered the *noncommutative unit ball* $\mathcal{B}_d$, that is, the set of all d-tuples of matrices $(X_1, \ldots, X_d)$ of arbitrary dimensions so that $||\sum_{i=1}^d X_i X_i^*|| < 1$, and they showed that the algebra

$$\mathrm{H}^\infty(\mathcal{B}_d) := \left\{ f \quad \text{nc function} \;\middle|\; \sup_{X\in\mathcal{B}_d} ||f(X)|| < \infty \right\}$$

is the multiplier algebra of a suitable nc reproducing kernel Hilbert space, defined as the *nc Drury-Arveson space* $\mathrm{H}^2(\mathcal{B}_d)$. The nc kernel that defines $\mathrm{H}^2(\mathcal{B}_d)$ is called the *nc Szegö kernel*, and it has a noncommutative version of the complete Pick property, [8]. Definition 1 extends naturally to the nc setting as well: we say that a sequence $A = (A_n)_{n\in\mathbb{N}}$ in $\mathcal{B}_d$ is interpolating if given any bounded sequence $(\phi_n)_{n\in\mathbb{N}}$ in $\mathrm{H}^\infty(\mathcal{B}_d)$ there exists a function ϕ in $\mathrm{H}^\infty(\mathcal{B}_d)$ so that $\phi(A_n) = \phi_n(A_n)$, for any n in $\mathbb{N}$. The second main result of this talk is a characterization of interpolating sequences in this nc setting, stated in terms of separation conditions in the nc Drury Arveson space $\mathrm{H}^2(\mathcal{B}_d)$:

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Theorem 2. *A is interpolating if and only if the sequence of subspaces*

$$\mathcal{H}_n := \{f \in \mathrm{H}^2(\mathcal{B}_d) | f(A_n) = 0\}^{\perp} \tag{1}$$

is a Riesz system.

In order to extend all points of Theorem 1 to this NC setting, it would be interesting to determine whether Carleson's interpolation theorems, [3] and [4], extend to the nc unit ball:

Question 3. *Let A be a sequence of d-tuples of matrices in $\mathcal{B}_d$. Is A interpolating, provided that the associated sequence $(\mathcal{H}_n)_{n\in\mathbb{N}}$ defined in* (1) *is strongly separated?*

Question 4. *Let A be a sequence of d-tuples of matrices in $\mathcal{B}_d$. Is A interpolating, provided that the associated sequence $(\mathcal{H}_n)_{n\in\mathbb{N}}$ defined in* (1) *is a weakly separated Bessel system?*

References

[1] Agler, J. and M$^{\mathrm{c}}$Carthy, J. E.: *Pick Interpolation and Hilbert Function Spaces*, Graduate Studies in Mathematics, volume 44, American Mathematical Society.
[2] Ball, J. A., Marx, G. and Vinnikov, V.: Noncommutative Reproducing Kernel Hilbert Spaces, *Journal of Functional Analysis*, Vol 271, Issue 7, 1, (2016), 1844-1920.
[3] Carleson, L.: An Interpolation Problem for Bounded Analytic Functions, *American Journal of Mathematics*, Vol. 80, No. 4 (Oct. 1958), pp. 921-930.
[4] Carleson, L.: Interpolation by Bounded Analytic Functions and the Corona Problem, *Annals if Mathematics*, Second Series, Vol. 76, No. 3 (Nov. 1962), pp. 547-559.
[5] Dayan, A.: Interpolating Matrices, *Integral Equations and Operator Theory* 92, 49 (2020)
[6] Dayan, A.: Interpolating d-tuples of Matrices, `arXiv 2106.00636`
[7] Dayan, A.: Weakly Separated Bessel Systems of Model Spaces, to be appearing in *Canadian Math. Bulletin*
[8] Salomon, G., Shalit, O. M. and Shamovich, E.: Algebras of Bounded Noncommutative Analytic Functions on Subvarieties of the Noncommutative Unit Ball, *Trans. Amer. Math. Soc.*, 370(12):8639–8690, 2018.
[9] Shapiro, H. S. and Shields, A. L.: On some interpolation problems for analytic functions, *Amer. J. Math.* 83 (1961) 513-532.

Department of Mathematics
Norwegian University of Science and Technology,
Trondheim, Norway
Email address: `alberto.dayan@ntnu.no`

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Completing an Operator Matrix and the Free Joint Numerical Radius ⋆

Kennett L. Dela Rosa * **Hugo J. Woerdeman** **

* *Institute of Mathematics, University of the Philippines Diliman, Quezon City, NCR 1101, Philippines (e-mail: pldelarosa@up.edu.ph).*
** *Department of Mathematics, Drexel University, Philadelphia, PA 19104, USA (e-mail: hugo@math.drexel.edu)*

Abstract: Ando's classical characterization of the unit ball in the numerical radius norm was generalized by Farenick, Kavruk, and Paulsen using the free joint numerical radius of a tuple of Hilbert space operators $(X_1, \dots, X_m)$. In particular, the characterization leads to a positive definite completion problem. In this paper, we study various aspects of Ando's result in this generalized setting. Among other things, this leads to the study of finding a positive definite solution L to the equation

$$L = I + \sum_{j=1}^{m} \left[\left(L^{\frac{1}{2}} X_j^* L X_j L^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} + \left(L^{\frac{1}{2}} X_j L X_j^* L^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} \right],$$

which may be viewed as a fixed point equation. Once such a fixed point is identified, the desired positive definite completion is easily obtained. Along the way we derive other related results including basic properties of the free joint numerical radius and an easy way to determine the free joint numerical radius of a tuple of generalized permutations. Finally, we present some open problems.

Keywords: Free joint numerical radius, Matrix completion, Fixed point, Generalized permutation

1. EXTENDED ABSTRACT

For a bounded Hilbert space operator $X \in B(\mathcal{H})$, the *numerical radius* is defined by

$$w(X) = \sup\{|\langle Xh, h\rangle| : \ h \in \mathcal{H}, \ \|h\| = 1\}.$$

The numerical radius corresponds to the radius of the smallest circle centered at 0 that contains the *numerical range*

$$W(X) = \{\langle Xh, h\rangle : \ h \in \mathcal{H}, \ \|h\| = 1\}.$$

The well known characterization by Ando (1973) of operators whose numerical radius is at most 1 states that $w(X) \leq 1$ if and only if there exists $Z = Z^* \in B(\mathcal{H})$ so that

$$\begin{bmatrix} I - Z & X \\ X^* & I + Z \end{bmatrix} \geq 0,$$

where $T \geq 0$ is shorthand for T being a positive semidefinite operator. Equivalently, $w(X) \leq 1$ if and only if there exist $A_1, A_2 \in B(\mathcal{H})$ with $A_1 + A_2 = I$ so that

$$\begin{bmatrix} A_1 & X/2 \\ X^*/2 & A_2 \end{bmatrix} \geq 0. \tag{1}$$

One way to prove Ando's result is to observe that $w(X) \leq 1$ if and only if

$$Q(e^{i\theta}) = I - \mathrm{Re}(e^{i\theta} X) \geq 0, \ \text{for all } \theta \in [0, 2\pi],$$

and subsequently use Fejér-Riesz factorization

$$\begin{aligned} I - zX/2 - \overline{z}X^*/2 &= Q(z) \\ &= (P_0 + P_1 z)^*(P_0 + P_1 z), \ |z| = 1. \end{aligned}$$

Now $P_0^* P_0 + P_1^* P_1 = I$ and $P_0^* P_1 = -X/2$ and thus

$$0 \leq \begin{bmatrix} P_0^* \\ -P_1^* \end{bmatrix} [P_0 \ -P_1] =: \begin{bmatrix} A_1 & X/2 \\ X^*/2 & A_2 \end{bmatrix}$$

where $A_1 = P_0^* P_0$ and $A_2 = P_1^* P_1$ satisfy $A_1 + A_2 = I$.

There are different ways to find A_1 and A_2 so that (1) holds. In finite dimensions, one can find A_1 and A_2 numerically by using semidefinite programming, as a block matrix (1) is in the intersection of the cone of positive semidefinite matrices and the affine space

$$\left\{ \begin{bmatrix} I & X/2 \\ X^*/2 & 0 \end{bmatrix} + \begin{bmatrix} -Z & 0 \\ 0 & Z \end{bmatrix} : Z = Z^* \right\}.$$

Semidefinite programming is exactly designed to handle such a situation.

An alternative process to arrive at (1), which was used by Ando in his original paper, is to consider $Z_k = Z_k^*$ defined via $\langle Z_k h, h\rangle$

$$\inf_{h_1, \dots, h_k} \left\langle \begin{bmatrix} I & X/2 & \cdots & 0 \\ X^*/2 & I & \ddots & \vdots \\ \vdots & \ddots & \ddots & X/2 \\ 0 & \cdots & X^*/2 & I \end{bmatrix} \begin{bmatrix} h \\ h_1 \\ \vdots \\ h_k \end{bmatrix}, \begin{bmatrix} h \\ h_1 \\ \vdots \\ h_k \end{bmatrix} \right\rangle.$$

Then Z_k converges decreasingly to Z, say; and we obtain

⋆ The research of the first author was supported by UP Diliman's Ph.D. Incentive Award. The research of the second author was supported by Simons Foundation grant 355645 and National Science Foundation grant DMS 2000037.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$$\begin{bmatrix} I - Z & X/2 \\ X^*/2 & Z \end{bmatrix} \geq 0 \tag{2}$$

yielding representation (1). In fact, this process yields the maximal Z in (2) (and gives a co-outer factorization of $Q(z)$). In the case when $w(X) < 1$, this leads to the iterative scheme

$$Z_1 = I \text{ and } Z_{k+1} = I - (X/2)Z_k^{-1}(X^*/2) \text{ for } k \in \mathbb{N},$$

which monotonically decreases; see Algorithm 4.1 in Engwerda et al. (1993).

Farenick et al. (2013) generalized Ando's result to the multivariable setting as follows.

Theorem 1. (Farenick et al., 2013, Theorem 3.4) Let $X_1, \ldots, X_m \in B(\mathcal{H})$. The following are equivalent:

(i) $w(X_1, \ldots, X_m) < 1/2$.
(ii) There exist $A_1, \ldots, A_{m+1} \in B(\mathcal{H})$ so that $A_1 + \cdots + A_{m+1} = I$ and

$$\begin{bmatrix} A_1 & X_1 & 0 & \cdots & 0 \\ X_1^* & A_2 & X_2 & & \vdots \\ 0 & X_2^* & \ddots & \ddots & 0 \\ \vdots & & \ddots & A_m & X_m \\ 0 & \cdots & 0 & X_m^* & A_{m+1} \end{bmatrix} > 0. \tag{3}$$

In (3), $T > 0$ is shorthand for T being a positive definite operator. Condition (i) in Theorem 1 concerns the *free joint numerical radius* of a tuple of m Hilbert space operators $X_1, \ldots, X_m \in B(\mathcal{H})$, defined as

$$w(X_1, \ldots, X_m) = \sup \{ w (X_1 \otimes U_1 + \cdots + X_m \otimes U_m) \},$$

where the supremum is taken over every Hilbert space $\mathcal{K}$, every choice of m unitaries $U_1, \ldots, U_m \in B(\mathcal{K})$, and the tensor product is *spatial*, which can be defined as follows. Consider an inner product on the algebraic tensor of $\mathcal{H}$ and $\mathcal{K}$ by letting $\langle h_1 \otimes k_1, h_2 \otimes k_2 \rangle := \langle h_1, h_2 \rangle_{\mathcal{H}} \cdot \langle k_1, k_2 \rangle_{\mathcal{K}}$ for all $h_1, h_2 \in \mathcal{H}$, $k_1, k_2 \in \mathcal{K}$, and then extending linearly. Denote by $\mathcal{H} \otimes \mathcal{K}$ the resulting Hilbert space after completion. For $R \in B(\mathcal{H})$ and $S \in B(\mathcal{K})$, consider defining a map $(R \otimes S)(h \otimes k) := (Rh) \otimes (Sk)$ for all $h \in \mathcal{H}$ and $k \in \mathcal{K}$, and then extending linearly. The resulting operator $R \otimes S$ has the property that $\|R \otimes S\| = \|R\| \cdot \|S\|$. Hence, the algebraic tensor of $B(\mathcal{H})$ and $B(\mathcal{K})$ naturally inherits a norm (called the *spatial tensor norm*) as a subset of $B(\mathcal{H} \otimes \mathcal{K})$. Taking the closure with respect to the spatial tensor norm yields a C^*-subalgebra of $B(\mathcal{H} \otimes \mathcal{K})$.

The free joint numerical radius coincides with the classical numerical radius when there is only one operator ($m = 1$), and Theorem 1 reduces to Ando's classical result. The objective of this paper is to pursue the different aspects of Ando's result in this more general setting. This includes (i) finding a solution using semidefinite programming; (ii) finding a solution via an iterative scheme (which may have the potential to generalize to the infinite dimensional case); and (iii) exploring the connection with factorization. As we will see, along the way we derive other related results including basic properties of the free joint numerical radius and an easy way to determine the free joint numerical radius of a tuple of generalized permutations.

Our approach to solve for $A_1, \ldots, A_{m+1}$ in (3) will be different than Ando's. We will show, in finite dimensions, that a solution $A_1, \ldots, A_{m+1}$ in (3) exists exactly when the function $f_{X_1,\ldots,X_m}$ defined below has a positive definite fixed point. For a given tuple $X_1, \ldots, X_m \in B(\mathcal{H})$ and for any $Z \geq 0$, define $f_{X_1,\ldots,X_m}(Z)$ as

$$I + \sum_{j=1}^{m} \left[\left(Z^{\frac{1}{2}} X_j^* Z X_j Z^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} + \left(Z^{\frac{1}{2}} X_j Z X_j^* Z^{\frac{1}{2}} + \frac{1}{4} I \right)^{\frac{1}{2}} \right]. \tag{4}$$

Operator monotonicity of $t^{\frac{1}{2}}$ implies

$$f_{X_1,\ldots,X_m}(Z) \geq (m+1)I > 0 \text{ for any } Z \geq 0.$$

Theorem 2. Let $X_1, \ldots, X_m \in B(\mathcal{H})$. Consider the following statements:

(i) There exists positive definite $L \in B(\mathcal{H})$ for which

$$f_{X_1,\ldots,X_m}(L) = L. \tag{5}$$

(ii) $w(X_1, \ldots, X_m) < \frac{1}{2}$.

Then (i)→(ii). If $\dim(\mathcal{H}) < \infty$, then (ii)→(i).

Corollary 3. Let $X_1, \ldots, X_m \in B(\mathcal{H})$ with $\dim(\mathcal{H}) < \infty$. Then (i) and (ii) in Theorem 2 are equivalent.

We prove Theorem 2 using matrix completion techniques. We will discuss the difficulties encountered in generalizing $(ii) \to (i)$ to the infinite dimensional case. Once a positive definite fixed point for $f_{X_1,\ldots,X_m}$ is identified, we show that there is an easy construction for the unknowns $A_1, \ldots, A_{m+1}$ in (3) (which works in all dimensions).

In order to find a solution L to (5), one can use well known iterative schemes to find such a fixed point, with the iterative scheme $L_{k+1} = f_{X_1,\ldots,X_m}(L_k)$ being the standard choice. The choice of a starting point is of course important, and we have found that the choice $L_1 = (m+1)I$ (which is the fixed point when $X_1 = \cdots = X_m = 0$) works perfectly numerically, and in fact we find that the corresponding sequence $\{L_k\}_{k \in \mathbb{N}}$ is monotonically nondecreasing in the Loewner partial ordering. Recall that the Loewner partial ordering on Hermitian operators is given by $R \leq S$ if and only if $S - R \geq 0$. This leads to the following conjecture.

Conjecture 4. Let $X_1, \ldots, X_m \in B(\mathcal{H})$. Consider the recurrence

$$L_1 = (m+1)I \text{ and } L_{k+1} = f_{X_1,\ldots,X_m}(L_k) \text{ for } k \in \mathbb{N}, \tag{6}$$

where $f_{X_1,\ldots,X_m}$ is defined in (4). Then

(i) $L_k \leq L_{k+1}$ for all $k \in \mathbb{N}$.
(ii) If $w(X_1, \ldots, X_m) < 1/2$, then $\{L_k\}_{k \in \mathbb{N}}$ converges in the weak operator topology to a fixed point $L \in B(\mathcal{H})$ of $f_{X_1,\ldots,X_m}$.

In general, $L_1 = (m+1)I \leq f(L_1) = L_2$. We will prove Conjecture 4 in the case when $X_1, \ldots, X_m$ are generalized permutations, i.e., each X_j is the product of a permutation matrix and a diagonal matrix. It is worthwhile to observe that our iterative scheme has a different origin than the iterative scheme from Ando's work. Indeed, in Ando's approach one maximizes A_2 in (1) (in the Loewner partial order) while our approach is based on maximizing the determinant of (1). Even though our approach is based on finite dimensional considerations, the iteration scheme can also be defined in infinite dimensional settings. It is our hope that a convergence proof for that case can be obtained in the future.

Aside from the results mentioned above, we will also cover the following. We will show some basic properties

of the free joint numerical radius. We will prove a closed formula for the free joint numerical radius of a tuple of n-by-n generalized permutations. We will describe how to use semidefinite programming to numerically compute $w(X_1, \ldots, X_m)$ for a tuple of n-by-n matrices. We will prove a limit formula for the free joint numerical radius of a tuple of generalized permutations on infinite dimensional separable Hilbert spaces. We will discuss the connection with factorization of Hermitian pencils.

REFERENCES

Ando, T. (1973). Structure of operators with numerical radius one. *Acta Sci. Math. (Szeged)*, 34, 11–15.

Engwerda, J.C., Ran, A.C.M., and Rijkeboer, A.L. (1993). Necessary and Sufficient Conditions for the Existence of a Positive Definite Solution of the Matrix Equation $X + A^*X^{-1}A = Q$. *Linear Algebra Appl.*, 186, 255–275.

Farenick, D., Kavruk, A.S., and Paulsen, V.I. (2013). C*-algebras with the weak expectation property and a multivariable analogue of Ando's theorem on the numerical radius. *J. Operator Theory*, 70, 573–590.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Structure-preserving integrators for dissipative systems based on reversible-irreversible splitting

Xiaocheng Shang * **Hans Christian Öttinger** **

* *School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom (e-mail: x.shang.1@bham.ac.uk).*
** *Department of Materials, Polymer Physics, ETH Zürich, Leopold-Ruzicka-Weg 4, CH-8093 Zürich, Switzerland (e-mail: hco@mat.ethz.ch)*

Abstract: We study the optimal design of numerical integrators for dissipative systems, for which there exists an underlying thermodynamic structure known as GENERIC (General Equation for the NonEquilibrium Reversible-Irreversible Coupling). We present a frame-work to construct structure-preserving integrators by splitting the system into reversible and irreversible dynamics. The reversible part, which is often degenerate and reduces to a Hamiltonian form on its symplectic leaves, is solved by using a symplectic method (e.g., Verlet) with degenerate variables being left unchanged, for which an associated modified Hamiltonian (and subsequently a modified energy) in the form of a series expansion can be obtained by using backward error analysis. The modified energy is then used to construct a modified friction matrix associated with the irreversible part in such a way that a modified degeneracy condition is satisfied. The modified irreversible dynamics can be further solved by an explicit midpoint method if not exactly solvable. Our findings are verified by various numerical experiments, demonstrating the superiority of structure-preserving integrators over alternative schemes in terms of not only the accuracy control of both energy conservation and entropy production but also the preservation of the conformal symplectic structure in the case of linearly damped systems.

Keywords: structure-preserving integrators, dissipative systems, GENERIC, conformal symplectic, discrete gradient methods.

1. INTRODUCTION

In the last few decades, considerable effort has been devoted to developing structure-preserving integrators for Hamiltonian systems. It has been demonstrated that the so-called symplectic integrators, which preserve the symplectic structure, have superior long time behavior compared to their nonsymplectic counterparts, and should be preferred in practice Hairer et al. (2006); Leimkuhler and Matthews (2015); Leimkuhler and Reich (2005). On the other hand, there has been growing interest in designing appropriate numerical methods for gradient flows Ambrosio et al. (2008); Hairer and Lubich (2014); Jordan et al. (1998); Otto (2001); Stuart and Humphries (1996) that respect their underlying properties. In contrast to the symplectic structure, the conformal symplectic structure Bhatt et al. (2016); Bhatt and Moore (2017); Dressler (1988); Hong et al. (2017); McLachlan and Perlmutter (2001); Moser (1994) for Hamiltonian systems that are perturbed by a linear damping (which can be thought of as a special case of the Rayleigh dissipation) has been less studied. It is also worth mentioning that variational integrators Kane et al. (2000) and specialized Runge–Kutta methods Jay (2003) have also been used to solve dissipative systems. It turns out that thermodynamically admissible evolution equations for nonequilibrium systems have a more general (including an additional variable known as entropy) and well-defined structure known as GENERIC (General Equation for the NonEquilibrium Reversible-Irreversible Coupling) Grmela and Öttinger (1997); Öttinger (2005, 2018); Öttinger and Grmela (1997), which possesses the following distinct features:

(i) conservation of the total energy;
(ii) separation of the reversible and irreversible dynamics;
(iii) the reversible dynamics preserves a Poisson structure;
(iv) entropy production is unaffected by the reversible dynamics;
(v) nonnegative entropy production rate.

More specifically, the GENERIC formulation of the time evolution for nonequilibrium systems is given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = L\frac{\partial E}{\partial x} + M\frac{\partial S}{\partial x}\,, \tag{1}$$

where x is the set of independent variables required to describe a given nonequilibrium system, E and S represent respectively the total energy and entropy as functions of

⋆ This is a resubmission of an extended abstract that was accepted for presentation at the MTNS 2020 in Cambridge.

the independent variables x, and L and M denote the antisymmetric Poisson matrix and the positive semidefinite (symmetric) friction matrix, respectively. Note that both L and M can also depend on the independent variables x so that the fundamental time evolution equation (1) could be highly nonlinear. We also point out that $\partial/\partial x$ in (1) simply implies the partial derivative although it typically denotes the functional derivative when x is a function/field. Moreover, (1) is supplemented by two degeneracy conditions:

$$L\frac{\partial S}{\partial x} = 0\,, \tag{2}$$

and

$$M\frac{\partial E}{\partial x} = 0\,. \tag{3}$$

Eqs. (2)–(3) indicate the conservation of the entropy by the reversible dynamics (i.e., the L contribution) and the conservation of the total energy in a closed system by the irreversible dynamics (i.e., the M contribution), respectively. Note that "reversible" and "irreversible" dynamics (in thermodynamics) are simply the names of the two fundamental contributions to the time evolution equation (1), and should not be confused with similar terms in other subjects. The rank of M has the interpretation of the number of dissipative processes taking place in the system. (See more discussions on the formulation of the GENERIC framework in Grmela and Öttinger (1997); Öttinger (2005, 2018); Öttinger and Grmela (1997).)

The usefulness and maturity of the GENERIC framework have been illustrated in a very large number of successful applications in a wide range of areas in Appendix E of Öttinger (2005) (see also a most recent review of Öttinger (2017) and references therein). In particular, despite its simple form, we believe that the irreversible dynamics in (1) is the most general form of meaningful irreversible equations in nonequilibrium thermodynamics—it is a belief based on both a very large variety of successful examples and statistical mechanics, so that it can be called knowledge (in particular, as this belief is widely accepted in the nonequilibrium thermodynamics community).

In order to further demonstrate the general properties of L and M, the Poisson bracket is given by

$$\{\mathcal{A},\mathcal{B}\} = \frac{\partial \mathcal{A}}{\partial x}\cdot L\frac{\partial \mathcal{B}}{\partial x}\,, \tag{4}$$

where $\mathcal{A}$ and $\mathcal{B}$ are sufficiently regular (and real-valued) functions of the independent variables x, and the dissipative bracket is given by

$$[\mathcal{A},\mathcal{B}] = \frac{\partial \mathcal{A}}{\partial x}\cdot M\frac{\partial \mathcal{B}}{\partial x}\,. \tag{5}$$

With the help of the two brackets and the chain rule, the time evolution equation of an arbitrary function $\mathcal{A}$ can then be written as

$$\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}t} = \{\mathcal{A},E\} + [\mathcal{A},S]\,. \tag{6}$$

More specifically, the Poisson bracket (4) inherits the antisymmetry of L,

$$\{\mathcal{A},\mathcal{B}\} = -\{\mathcal{B},\mathcal{A}\}\,, \tag{7}$$

and satisfies the Leibniz rule,

$$\{\mathcal{A}\mathcal{B},\mathcal{C}\} = \mathcal{A}\{\mathcal{B},\mathcal{C}\} + \mathcal{B}\{\mathcal{A},\mathcal{C}\}\,, \tag{8}$$

where $\mathcal{C}$ is another arbitrary sufficiently regular (and real-valued) function of the independent variables x. In addition, the Poisson bracket is required to satisfy the Jacobi identity,

$$\{\mathcal{A},\{\mathcal{B},\mathcal{C}\}\} + \{\mathcal{B},\{\mathcal{C},\mathcal{A}\}\} + \{\mathcal{C},\{\mathcal{A},\mathcal{B}\}\} = 0\,. \tag{9}$$

The dissipative bracket (5) inherits the symmetry of M,

$$[\mathcal{A},\mathcal{B}] = [\mathcal{B},\mathcal{A}]\,, \tag{10}$$

and also satisfies the Leibniz rule,

$$[\mathcal{A}\mathcal{B},\mathcal{C}] = \mathcal{A}[\mathcal{B},\mathcal{C}] + \mathcal{B}[\mathcal{A},\mathcal{C}]\,. \tag{11}$$

The positive semidefinite nature of M leads to the nonnegativeness condition

$$[\mathcal{A},\mathcal{A}] \geq 0\,, \tag{12}$$

which implies the second law of nonequilibrium thermodynamics (i.e., the entropy production rate is always nonnegative),

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\partial S}{\partial x}\cdot M\frac{\partial S}{\partial x} = [S,S] \geq 0\,. \tag{13}$$

This article addresses the long-standing challenge of how to preserve the underlying structures when numerically discretizing GENERIC systems in practice. Although in recent years this topic has attracted increasing attention Kraus and Hirvijoki (2017); Krüger et al. (2016, 2011); Morrison (2017); Portillo et al. (2017), to the best of our knowledge, there are no such numerical integrators in the literature. Unlike common approaches that are based on exact energy conservation, we propose in this article a framework to construct structure-preserving integrators for dissipative systems, i.e., GENERIC integrators (also known as metriplectic integrators Grmela (1984); Kaufman (1984); Morrison (1984, 1986) in the mathematical literature), based on splitting the reversible and irreversible dynamics. The topic of structure-preserving integrators for GENERIC/metriplectic systems is the counterpart and generalization of the theory of symplectic integrators for Hamiltonian systems.

1.1 Full GENERIC integrators

We recall the definition of (full) GENERIC integrators given in Öttinger (2018). Analogous to the definition of symplectic integrators for Hamiltonian dynamics Moser (1968), a mapping, $x_0 \mapsto x_h$, is said to be a full GENERIC integrator if it corresponds to a continuous time evolution of a modified GENERIC system

$$\frac{\mathrm{d}x}{\mathrm{d}t} = L\frac{\partial \tilde{E}_h}{\partial x} + \tilde{M}_h\frac{\partial S}{\partial x}\,, \tag{14}$$

where $\tilde{E}_h$ and $\tilde{M}_h$ represent the modified energy and friction matrix associated with the integrator, respectively, satisfying a modified degeneracy condition:

$$\tilde{M}_h\frac{\partial \tilde{E}_h}{\partial x} = 0\,. \tag{15}$$

That is, given initial conditions $x(0) = x_0$, the analytical solution of (14), $x(t)$, should agree with what we obtain

from the integrator at time h, i.e., $x(h) = x_h$. A full GENERIC integrator $x \mapsto x_h$, which can be thought of as the formal solution of (14), possesses the following structure:

$$x_h = \exp\left\{ h \left(L\frac{\partial \tilde{E}_h}{\partial x} + \tilde{M}_h \frac{\partial S}{\partial x} \right) \cdot \frac{\partial}{\partial x} \right\} x \,. \tag{16}$$

Similar to symplectic integrators for Hamiltonian dynamics, the modified energy, $\tilde{E}_h$, is strictly conserved by a GENERIC integrator. The physical energy E is expected to remain close to the modified energy, $\tilde{E}_h$, even for long integration periods. Additionally, the modified friction matrix, $\tilde{M}_h$, should not introduce any additional dissipative processes not present in the original matrix M. We point out that full GENERIC integrators may only be available in special cases, for instance, a full GENERIC integrator in the case of a damped harmonic oscillator, where analytical solutions of the GENERIC system can be obtained, was proposed and discussed in Öttinger (2018). However, it should be noted that it is highly unlikely that analytical solutions would be available for general GENERIC systems. (Nevertheless, it might be eventually possible to recognise a full GENERIC integrator without exact solutions.) Therefore, in what follows we introduce a framework to construct "split" GENERIC integrators.

1.2 Split GENERIC integrators

Inspired by recent developments on splitting methods Abdulle et al. (2015); Leimkuhler et al. (2016); Leimkuhler and Matthews (2015, 2013a,b); Leimkuhler and Shang (2015, 2016a,b); Shang et al. (2017), we consider to split the reversible and irreversible parts of the GENERIC system in such a way that the reversible dynamics, which is often degenerate but possesses a Hamiltonian form on its symplectic leaves, can be integrated by using a symplectic method (e.g., Verlet) with degenerate variables being left unchanged, while the irreversible part (gradient flow) can be solved in such a way that as many structure elements as possible can be preserved (see more references on the challenging task of structure preservation on manifolds in Ambrosio et al. (2008); Hairer and Lubich (2014); Jordan et al. (1998); Matthes and Plazotta (2019); Matthes and Osberger (2014); Otto (2001); Stuart and Humphries (1996)).

An interesting question for the split GENERIC integrators is: under what conditions do a modified energy and an associated friction matrix, satisfying the modified degeneracy condition (15), exist? If they exist, how much do we know about their respective forms? GENERIC integrators share some common features of GENERIC systems discussed at the beginning of this article, which can also be thought of as the requirements for GENERIC integrators. Denoting the Jacobian matrix of the independent variables x as Ω, we have

(i) preservation of the Poisson structure for the reversible dynamics: $\Omega(x_0)L(x_0)\Omega^{\mathrm{T}}(x_0) = L(x_h)$;
(ii) nonnegative entropy production rate: $S(x_h) \geq S(x_0)$;
(iii) the modified degeneracy condition (15) is satisfied with the other (2) being unchanged;
(iv) preservation of the rank of the friction matrix: $\mathrm{rank}(\tilde{M}_h) = \mathrm{rank}(M)$.

As pointed out in Quispel and McLaren (2008), it has been proved in Zhong and Marsden (1988) that there cannot exist an integrator for "non-integrable" Hamiltonian dynamics that preserves both the symplectic (Poisson) structure and the energy (Hamiltonian). In fact, it has been discussed in Simo et al. (1992) that the preservation of either property has its advantages and disadvantages. While previous attempts to construct structure-preserving integrators for dissipative systems have been relying on the exact conservation of energy (i.e., the energy-conserving discrete gradient methods Cohen and Hairer (2011); McLachlan et al. (1999); Quispel and Turner (1996)), there is no obvious reason why integrators that preserve the Poisson structure for the reversible dynamics should be ignored.

REFERENCES

Abdulle, A., Vilmart, G., and Zygalakis, K.C. (2015). Long time accuracy of Lie–Trotter splitting methods for Langevin dynamics. *SIAM J. Numer. Anal.*, 53(1), 1–16. doi:10.1137/140962644. URL `http://dx.doi.org/10.1137/140962644`.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media.

Bhatt, A., Floyd, D., and Moore, B.E. (2016). Second order conformal symplectic schemes for damped Hamiltonian systems. *J. Sci. Comput.*, 66(3), 1234–1259.

Bhatt, A. and Moore, B.E. (2017). Structure-preserving exponential Runge–Kutta methods. *SIAM J. Sci. Comput.*, 39(2), A593–A612.

Cohen, D. and Hairer, E. (2011). Linear energy-preserving integrators for Poisson systems. *BIT*, 51(1), 91–101. doi:10.1007/s10543-011-0310-z. URL `https://doi.org/10.1007/s10543-011-0310-z`.

Dressler, U. (1988). Symmetry property of the Lyapunov spectra of a class of dissipative dynamical systems with viscous damping. *Phys. Rev. A*, 38(4), 2103.

Grmela, M. (1984). Bracket formulation of dissipative fluid mechanics equations. *Phys. Lett. A*, 102(8), 355–358.

Grmela, M. and Öttinger, H.C. (1997). Dynamics and thermodynamics of complex fluids. I. Development of a general formalism. *Phys. Rev. E*, 56(6), 6620.

Hairer, E. and Lubich, C. (2014). Energy-diminishing integration of gradient systems. *IMA J. Numer. Anal.*, 34(2), 452–461.

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer.

Hong, J., Sun, L., and Wang, X. (2017). High order conformal symplectic and ergodic schemes for the stochastic Langevin equation via generating functions. *SIAM J. Numer. Anal.*, 55(6), 3006–3029.

Jay, L.O. (2003). Solution of index 2 implicit differential-algebraic equations by Lobatto Runge–Kutta methods. *BIT*, 43(1), 93–106.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1), 1–17.

Kane, C., Marsden, J.E., Ortiz, M., and West, M. (2000). Variational integrators and the Newmark algorithm for

conservative and dissipative mechanical systems. *Int. J. Numer. Meth. Engng*, 49(10), 1295–1325.

Kaufman, A.N. (1984). Dissipative Hamiltonian systems: A unifying principle. *Phys. Lett. A*, 100(8), 419–422.

Kraus, M. and Hirvijoki, E. (2017). Metriplectic integrators for the Landau collision operator. *Phys. Plasmas*, 24(10), 102311.

Krüger, M., Groß, M., and Betsch, P. (2016). An energy-entropy-consistent time stepping scheme for nonlinear thermo-viscoelastic continua. *ZAMM Z. Angew. Math. Mech.*, 96(2), 141–178.

Krüger, M., Groß, M., and Betsch, P. (2011). A comparison of structure-preserving integrators for discrete thermoelastic systems. *Comput. Mech.*, 47(6), 701–722.

Leimkuhler, B., Matthews, C., and Stoltz, G. (2016). The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA J. Numer. Anal.*, 36(1), 13–79. doi:10.1093/imanum/dru056. URL `http://dx.doi.org/10.1093/imanum/dru056`.

Leimkuhler, B. and Matthews, C. (2015). *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Springer.

Leimkuhler, B. and Matthews, C. (2013a). Rational construction of stochastic numerical methods for molecular sampling. *Appl. Math. Res. Express*, 2013(1), 34–56. doi:10.1093/amrx/abs010. URL `http://dx.doi.org/10.1093/amrx/abs010`.

Leimkuhler, B. and Matthews, C. (2013b). Robust and efficient configurational molecular sampling via Langevin dynamics. *J. Chem. Phys.*, 138, 174102. doi:10.1063/1.4802990. URL `http://dx.doi.org/10.1063/1.4802990`.

Leimkuhler, B. and Reich, S. (2005). *Simulating Hamiltonian Dynamics*. Cambridge University Press.

Leimkuhler, B. and Shang, X. (2015). On the numerical treatment of dissipative particle dynamics and related systems. *J. Comput. Phys.*, 280, 72–95. doi:10.1016/j.jcp.2014.09.008. URL `http://dx.doi.org/10.1016/j.jcp.2014.09.008`.

Leimkuhler, B. and Shang, X. (2016a). Adaptive thermostats for noisy gradient systems. *SIAM J. Sci. Comput.*, 38(2), A712–A736. doi:10.1137/15M102318X. URL `http://dx.doi.org/10.1137/15M102318X`.

Leimkuhler, B. and Shang, X. (2016b). Pairwise adaptive thermostats for improved accuracy and stability in dissipative particle dynamics. *J. Comput. Phys.*, 324, 174–193. doi:10.1016/j.jcp.2016.07.034. URL `http://dx.doi.org/10.1016/j.jcp.2016.07.034`.

Matthes, D. and Plazotta, S. (2019). A variational formulation of the BDF2 method for metric gradient flows. *ESAIM Math. Model. Numer. Anal.*, 53(1), 145–172. doi:10.1051/m2an/2018045. URL `https://doi.org/10.1051/m2an/2018045`.

Matthes, D. and Osberger, H. (2014). Convergence of a variational Lagrangian scheme for a nonlinear drift diffusion equation. *ESAIM Math. Model. Numer. Anal.*, 48(3), 697–726. doi:10.1051/m2an/2013126. URL `https://doi.org/10.1051/m2an/2013126`.

McLachlan, R. and Perlmutter, M. (2001). Conformal Hamiltonian systems. *J. Geom. Phys.*, 39(4), 276–300.

McLachlan, R.I., Quispel, G.R.W., and Robidoux, N. (1999). Geometric integration using discrete gradients. *Phil. Trans. R. Soc. A*, 357(1754), 1021–1045. doi:10.1098/rsta.1999.0363. URL `https://doi.org/10.1098/rsta.1999.0363`.

Morrison, P.J. (1984). Bracket formulation for irreversible classical fields. *Phys. Lett. A*, 100(8), 423–427.

Morrison, P.J. (1986). A paradigm for joined Hamiltonian and dissipative systems. *Physica D*, 18(1-3), 410–419.

Morrison, P.J. (2017). Structure and structure-preserving algorithms for plasma physics. *Phys. Plasmas*, 24(5), 055502.

Moser, J. (1968). Lectures on Hamiltonian systems. *Mem. Amer. Math. Soc.*, (81), 1–60.

Moser, J. (1994). On quadratic symplectic mappings. *Math. Z.*, 216(1), 417–430.

Öttinger, H.C. (2005). *Beyond Equilibrium Thermodynamics*. John Wiley & Sons.

Öttinger, H.C. (2017). GENERIC: Review of successful applications and a challenger for the future. In *14th Joint European Thermodynamics Conference*.

Öttinger, H.C. (2018). GENERIC integrators: Structure preserving time integration for thermodynamic systems. *J. Non-Equilib. Thermodyn.*, 43(2), 89–100.

Öttinger, H.C. and Grmela, M. (1997). Dynamics and thermodynamics of complex fluids. II. Illustrations of a general formalism. *Phys. Rev. E*, 56(6), 6633.

Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2), 101–174.

Portillo, D., García Orden, J.C., and Romero, I. (2017). Energy–entropy–momentum integration schemes for general discrete non-smooth dissipative problems in thermomechanics. *Int. J. Numer. Meth. Engng*.

Quispel, G.R.W. and McLaren, D.I. (2008). A new class of energy-preserving numerical integration methods. *J. Phys. A: Math. Theor.*, 41(4), 045206. doi:10.1088/1751-8113/41/4/045206. URL `https://doi.org/10.1088/1751-8113/41/4/045206`.

Quispel, G.R.W. and Turner, G.S. (1996). Discrete gradient methods for solving ODEs numerically while preserving a first integral. *J. Phys. A: Math. Gen.*, 29(13), L341. doi:10.1088/0305-4470/29/13/006. URL `https://doi.org/10.1088/0305-4470/29/13/006`.

Shang, X., Kröger, M., and Leimkuhler, B. (2017). Assessing numerical methods for molecular and particle simulation. *Soft Matter*, 13, 8565–8578. doi:10.1039/C7SM01526G. URL `http://dx.doi.org/10.1039/C7SM01526G`.

Simo, J.C., Tarnow, N., and Wong, K.K. (1992). Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics. *Comput. Methods Appl. Mech. Eng.*, 100(1), 63–116. doi:10.1016/0045-7825(92)90115-Z. URL `https://doi.org/10.1016/0045-7825(92)90115-Z`.

Stuart, A.M. and Humphries, A.R. (1996). *Dynamical Systems and Numerical Analysis, Volume 2 of Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press.

Zhong, G. and Marsden, J.E. (1988). Lie–Poisson Hamilton–Jacobi theory and Lie–Poisson integrators. *Phys. Lett. A*, 133(3), 134–139. doi:10.1016/0375-9601(88)90773-6. URL `https://doi.org/10.1016/0375-9601(88)90773-6`.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Controller synthesis for an arbitrary length of mass chain

Kaoru Yamamoto *

* *Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan.* yamamoto@ees.kyushu-u.ac.jp.

Abstract: The disturbance suppression problem for a chain of masses is discussed. The particular focus is placed on synthesising mechanical networks between masses that effectively suppress the disturbance propagation along the chain of any length. This study is motivated by the problem of controlling multi-agent systems where agents may leave or join the network. That is, the size of the network may change over time. In this work, we give the explicit expressions of scalar transfer functions from disturbance to an intermass displacement as a function of the number of masses, N, and discuss the methodology of synthesising a controller such that the H^∞ norm is upper bounded by a prescribed value for any N.

Keywords: Mechanical networks, linear systems, decentralized systems, large-scale systems

1. INTRODUCTION

Decentralised control of multi-agent systems has been an active area of research with various applications such as consensus problems and flocking/formation/coverage control of mobile robots.

In such applications, one interesting problem is ensuring the ability to control a system behaviour even if the size of the system changes over time. This allows agents to join or leave the network in a flexible manner. In practice, however, such criteria are rarely satisfied since several key performance measures relating to global behaviours of multi-agent systems simply do not scale. Notable examples include the string instability (e.g., Seiler et al. (2004); Feng et al. (2019)) or network incoherence phenomena (e.g., Bamieh et al. (2012)). Nevertheless, it appears that average or local performance measures, for example those in Carli et al. (2009); Pates (2015); Bamieh et al. (2012); Yamamoto and Smith (2016); Pates and Yamamoto (2018), can be guaranteed independent of the size of the network.

Our focus here is to analyse how the agents react locally against disturbances in a multi-agent system modelled by a mass chain depicted in Fig. 1. Bidirectional control of a platoon of vehicles is one such example. In particular, we explicitly express the transfer function from the disturbance to a given intermass displacement as a function of the number of masses and propose a method to design a suitable controller that works well for any length of the chain.

This note is a resubmission of the work (Yamamoto (2021)) that has been accepted for presentation at the MTNS 2020 in Cambridge, but which was cancelled due to COVID-19.

General notation

The set of natural, real and complex numbers is denoted by $\mathbb{N}$, $\mathbb{R}$, $\mathbb{C}$, respectively. $\mathbb{R}^{m\times n}$ is the set of m-by-n real matrices. $\mathbb{C}_+$ is the closed right-half plane. H^∞ is the standard Hardy space on the right-half plane and $\|\cdot\|_\infty$ represents the H^∞-norm. The (i,j) entry of a matrix A is denoted by $[A]_{i,j}$. $\hat{f}$ denotes the Laplace transform of a signal f.

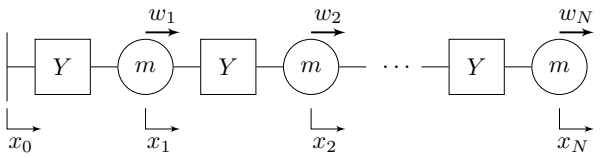


Fig. 1. Chain of N masses m connected by a mechanical admittance Y and connected to a movable point x_0. The disturbance on the ith mass is denoted as w_i.

2. CHAIN MODEL

We consider a chain of N identical point masses m connected by identical mechanical networks (Fig. 1). Each mechanical network provides an equal and opposite force on each mass and is assumed here to have negligible mass. It is also assumed that each mechanical network consists of a spring component in parallel with other components, i.e., the admittance takes the following form: $Y(s) = k/s + Y_c(s)$ where k is the spring coefficient. Here we assume that k is a given parameter and our task is to find a controller represented by an admittance $Y_c(s)$ that effectively supresses the disturbance in the chain.

The system is excited by a movable point $x_0(t)$ and external force acting on the ith mass, $w_i(t)$, $i \in \{1, 2, \ldots, N\}$. The displacement of the ith mass is then denoted by $x_i(t)$. Assuming that the system is initially at rest, the equations of motion in the Laplace transformed domain are given as

$$ms^2\hat{x} = sY(s)L_N\hat{x} + sY(s)\phi_1\hat{x}_0 + \hat{w},$$

and hence,

$$\hat{x} = (h(s)I - L_N)^{-1}\phi_1\hat{x}_0 + \frac{1}{sY(s)}(h(s)I - L_N)^{-1}\hat{w} \quad (1)$$

where I is the identity matrix,

$$h(s) \coloneqq ms/Y(s), \quad \hat{x} \coloneqq [\hat{x}_1, \dots, \hat{x}_N]^\top,$$
$$\hat{w} \coloneqq [\hat{w}_1, \dots, \hat{w}_N]^\top, \quad \phi_1 \coloneqq [1, 0, \dots, 0]^\top \in \mathbb{R}^N,$$
$$L_N \coloneqq \begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

The intermass displacement $e_i \coloneqq x_{i-1} - x_i$ is given by

$$\hat{e} = G_{ex_0}(s)\hat{x}_0 + G_{ew}(s)\hat{w} \tag{2}$$

where $\hat{e} = [\hat{e}_1, \dots, \hat{e}_N]^\top$ and

$$\begin{aligned} G_{ex_0}(s) &= \left(I + M(h(s)I - L_N)^{-1}\right)\phi_1 \\ &=: [G_{e_1x_0}(s), \dots, G_{e_Nx_0}(s)]^\top, \\ G_{ew}(s) &= \frac{1}{sY(s)} M(h(s)I - L_N)^{-1}, \end{aligned}$$
$$M = \begin{bmatrix} -1 & 0 & \cdots & \cdots & 0 \\ 1 & -1 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

To obtain an explicit inverse of the tridiagonal matrix $hI - L_N$, let us introduce the characteristic polynomials of $L_i \in \mathbb{R}^{i\times i}$ in the variable h:

$$d_i \coloneqq \det(hI - L_i),$$

and also the characteristic polynomials of $\overline{L}_i \in \mathbb{R}^{i\times i}$ in h:

$$\overline{d}_i \coloneqq \det(hI - \overline{L}_i)$$

where

$$\overline{L}_i \coloneqq \begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{bmatrix}.$$

Then $d_1 = h + 1$ and $\overline{d}_1 = h + 2$. Using the Laplace expansion, we find that

$$\begin{aligned} d_i(h) &= (h+2)d_{i-1}(h) - d_{i-2}(h), \\ \overline{d}_i(h) &= (h+2)\overline{d}_{i-1}(h) - \overline{d}_{i-2}(h), \quad \text{for } i = 1, \dots, N \end{aligned} \tag{3}$$

with initial conditions

$$d_{-1} = 1, d_0 = 1, \overline{d}_{-1} = 0, \overline{d}_0 = 1.$$

The inverse of $hI - L_N$ is then given as

$$\left[(hI - L_N)^{-1}\right]_{i,j} = \overline{d}_{j-1}(h)d_{N-i}(h) \quad \text{for } i \geq j, \tag{4}$$

which can be easily derived using the theorem provided by Usmani (1994). Note that $hI - L_N$ is symmetric and hence $\left[(hI - L_N)^{-1}\right]_{i,j} = \left[(hI - L_N)^{-1}\right]_{j,i}$.

The transfer functions $G_{e_ix_0}(s)$ and $G_{ew}(s)$ in (2) are then written as (suppressing the dependence on $h(s)$ in d_i and $\overline{d}_i$)

$$G_{e_ix_0}(s) = \frac{d_{N-i+1} - d_{N-i}}{d_N}, \tag{5}$$

$$G_{e_iw_j}(s) = \begin{cases} \dfrac{1}{sY(s)}\dfrac{1}{d_N}\overline{d}_{j-1}(d_{N-i+1} - d_{N-i}) & \text{for } i > j, \\ \dfrac{1}{sY(s)}\dfrac{1}{d_N}d_{N-j}(\overline{d}_{i-2} - \overline{d}_{i-1}) & \text{for } i \leq j, \end{cases} \tag{6}$$

where $G_{e_iw_j}(s)$ is the (i,j)-entry of $G_{ew}(s)$.

We say that the mass chain of Fig. 1 is stable if all poles in the transfer functions $G_{ex_0}(s)$ and $G_{e_iw_j}(s)$ have negative real parts. The following proposition gives a sufficient condition for the stability:

Proposition 1. The mass chain of Fig. 1 is stable if $h(s) \in \mathbb{C} \setminus (-4, 0)$ and $sY(s) \neq 0$ for all $s \in \mathbb{C}_+$.

Proof. See (Yamamoto, 2021, Theorem 4). □

3. INTERMASS DISPLACEMENTS

We now define the following functions:

$$F_N^{(i,j)}(h) := \begin{cases} \dfrac{1}{d_N}\overline{d}_{j-1}(d_{N-i+1} - d_{N-i}) & \text{for } i > j, \\ \dfrac{1}{d_N}d_{N-j}(\overline{d}_{i-2} - \overline{d}_{i-1}) & \text{for } i \leq j. \end{cases} \tag{7}$$

Then, for $h(s) = ms/Y(s)$,

$$G_{e_1x_0}(s) = 1 + F_N^{(1,1)}(h(s)), \tag{8}$$

$$G_{e_ix_0}(s) = F_N^{(i,1)}(h(s)) \text{ for } i > 1, \tag{9}$$

$$G_{e_iw_j}(s) = \frac{1}{sY(s)}F_N^{(i,j)}(h(s)). \tag{10}$$

Treating h as an independent variable, the following theorem gives these closed-form expressions of $F_N^{(i,j)}(h)$:

Theorem 2. Let $\zeta \in \mathbb{C}$ be the root of

$$z^2 - (h+2)z + 1 = 0$$

satisfying $|\zeta| \leq 1$. For any $i, j \in \mathbb{N}$,

$$F_N^{(i,j)}(h) = \begin{cases} \dfrac{\zeta^{i-j}(1-\zeta^{2j})\left(1-\zeta^{2(N-i+1)}\right)}{(1+\zeta)\left(1+\zeta^{2N+1}\right)} & \text{for } i > j, \\ -\dfrac{\zeta^{j-i+1}(1+\zeta^{2i-1})\left(1+\zeta^{2(N-j)+1}\right)}{(1+\zeta)\left(1+\zeta^{2N+1}\right)} & \text{for } i \leq j. \end{cases} \tag{11}$$

suppressing the dependence on h in ζ.

Proof. See (Yamamoto, 2021, Theorem 4). □

3.1 Limits of the Sequences

It may be observed that the sequence $\left(F_N^{(i,j)}\right)$ in (11) is convergent for a fixed $\zeta \in \mathbb{C}$ with $|\zeta| < 1$ or $\zeta = \pm 1$, and divergent otherwise. The following theorem provides the condition for the convergence and its limit.

Theorem 3. The sequence $\left(F_N^{(i,j)}\right)$ converges pointwise to a limit $\mu^{(i,j)}$ for each $h \in \mathbb{C} \setminus (-4, 0)$ but fails to converge otherwise. In particular,

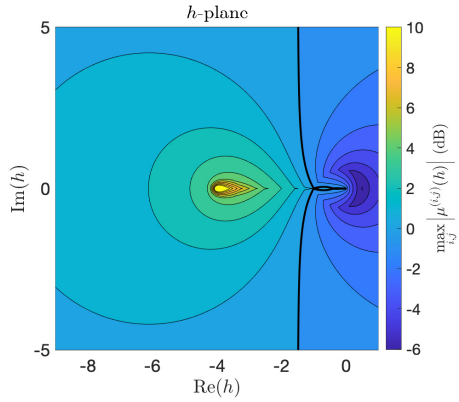


Fig. 2. Contour plot of $\max_{i,j} |\mu^{(i,j)}(h)|$ with $1 \le i, j \le 100$. The thick curve represents a contour of level 0 dB.

(1) For $h \in \mathbb{C} \setminus [-4, 0)$,

$$\mu^{(i,j)}(h) = \begin{cases} \dfrac{\zeta^{i-j}(1-\zeta^{2j})}{(1+\zeta)} & \text{for } i > j, \\ -\dfrac{\zeta^{j-i+1}(1+\zeta^{2i-1})}{(1+\zeta)} & \text{for } i \le j. \end{cases} \tag{12}$$

(2) For $h = -4$,

$$\mu^{(i,j)}(h) = \begin{cases} (-1)^{j-1} 2j & \text{for } i > j, \\ (-1)^{i-j-2}(2i-1) & \text{for } i \le j. \end{cases} \tag{13}$$

Proof. See (Yamamoto, 2021, Theorem 5). □

Hence, if $h(s) \in \mathbb{C} \setminus (-4, 0)$ for all $s \in \mathbb{C}_+$,

$$\sup_{\omega} \lim_{N\to\infty} |F_N^{(i,j)}(h(j\omega))| = \sup_{\omega} |\mu^{(i,j)}(h(j\omega))|. \tag{14}$$

Furthermore,

$$\sup_{N} \|F_N^{(i,j)}(h(s))\|_\infty \ge \sup_{\omega} |\mu^{(i,j)}(h(j\omega))|. \tag{15}$$

That is, $\sup_\omega |\mu^{(i,j)}(h(j\omega))|$ gives a lower bound of the supremum of H^∞-norm of $F_N^{(i,j)}(h(s))$ over N. A contour plot of the maximum magnitude of $\mu^{(i,j)}(h)$ over (i, j) with $0 \le i, j \le 100$ in the h-plane is shown in Fig. 2. The thick black curve represents $\max_{i,j} |\mu^{(i,j)}(h)| = 0$ (dB). The figure shows that the asymptotic value of $F_N^{(i,j)}(h(j\omega))$ as $N \to \infty$ is directly related to the proximity of $h(j\omega)$ to the point -4. From Theorem 3 we see that the magnitude of $|\mu^{(i,j)}(-4)|$ grows as we increase the indices and we must avoid this region.

3.2 Disturbance Amplification

In this subsection, we provide a graphical mean to design the interconnection admittance $Y(s)$.

Figure 3 shows a contour plot of $\max_{1\le N\le 200} |1 + F_N^{(1,1)}(h)|$ with the thick black curve representing a contour of level 0 dB. If the Nyquist diagram of $h(s)$ lies inside a contour of level γ, $\|G_{e_1x_0}(s)\|_\infty \le \gamma$ for any $N \in \mathbb{N}$. (Note that $G_{e_1x_0}(s) = 1 + F_N^{(1,1)}(h(s))$ from (8).) To demonstrate this, the Nyquist diagram of $h(s) = s^2/(2s+1)$ is also plotted in Fig. 3. Since it lies inside the curve of level 0 dB, $\|G_{e_1x_0}(s)\|_\infty \le 1$ for any $N \in \mathbb{N}$, as we can see in

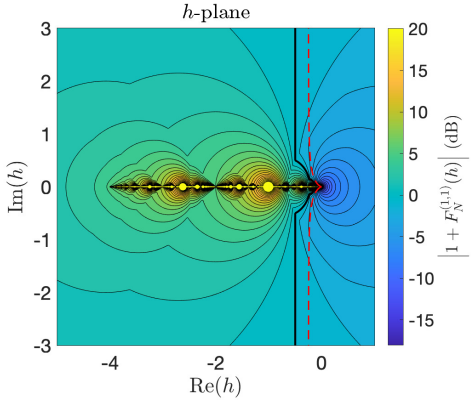


Fig. 3. Nyquist diagram of $h(s) = s^2/(2s+1)$ (red, dashed) and contour plot of $\max_{1\le N\le 200} |1+F_N^{(1,1)}|$. The thick black curve represents a contour of level 0 dB.

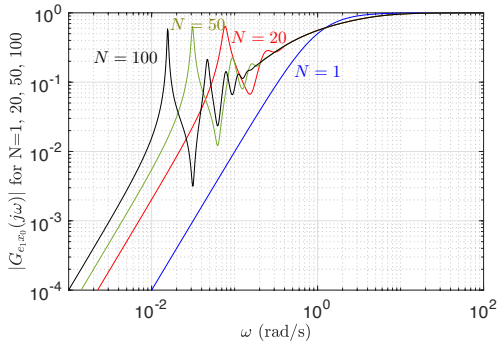


Fig. 4. The magnitude plot of $G_{e_1x_0}(j\omega)$ with $h(s) = s^2/(2s+1)$ for $N = 1, 20, 50, 100$.

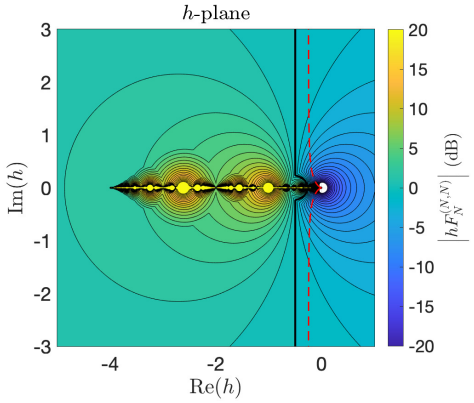


Fig. 5. Nyquist diagram of $h(s) = s^2/(2s+1)$ (red, dashed) and contour plot of $\max_{1\le N\le 200} |hF_N^{(1,N)}|$. The thick black curve represents a contour of level 0 dB.

Fig. 4 for $N = 1, 20, 50, 100$. $G_{e_ix_0}(s)$ can be evaluated similarly.

To evaluate $G_{e_iw_j}(s)$ in a similar way, we first rewrite (10) as

$$G_{e_iw_j}(s) = \frac{1}{sY(s)} F_N^{(i,j)}(h(s)) = \frac{1}{ms^2} h(s) F_N^{(i,j)}(h(s)) \tag{16}$$

and draw a contour plot of $\max_N |hF_N^{(i,j)}(h)|$. Figure. 5 shows this contour map for $(i, j) = (1, N)$ again with the

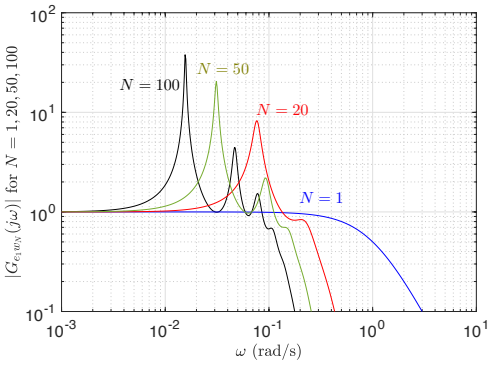


Fig. 6. The magnitude plot of $G_{e_1 w_N}(j\omega)$ with $h(s) = s^2/(2s+1)$ for $N = 1, 20, 50, 100$.

Nyquist diagram of $h(s) = s^2/(2s+1)$. Although Fig. 5 looks similar to Fig. 3, since $1/ms^2$ is multiplied as in (16), the same interconnection $h(s) = s^2/(2s+1)$ will result in large frequency response of $G_{e_1 w_N}(s)$ in the low frequency range. Indeed, this is observed in Fig. 6. To remedy this, we need to shape the locus of $h(j\omega)$ at $\omega \approx 0$ suitably. However, because of the existence of the parallel spring, the limiting behaviour of $h(j\omega)$ as $\omega \to 0$ cannot be drastically changed using passive interconnection, i.e., positive-real $Y(s)$. Whether the use of an active controller may lead to an improvement is yet to be explored.

4. CONCLUSION

Convenient representations of transfer functions from disturbance to a given intermass displacement in a homogeneous mass chain have been derived to evaluate how the system dynamics change as the number of masses N changes. The limiting behaviour of these transfer functions as N tends to infinity has been studied. Moreover, the possibility of designing a controller that achieves a prespecified disturbance attenuation level independent of N has been explored. Such a size-independent performance can be achieved using passive interconnection when only the movable point displacement is present as the disturbance. However, it is not the case when the disturbance on each mass is present. The use of active controllers is considered as a future work.

5. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP19H02161 and JP20K14766.

REFERENCES

Bamieh, B., Jovanovic, M., Mitra, P., and Patterson, S. (2012). Coherence in large-scale networks: Dimension-dependent limitations of local feedback. *IEEE Transactions on Automatic Control*, 57(9), 2235–2249.

Carli, R., Garin, F., and Zampieri, S. (2009). Quadratic indices for the analysis of consensus algorithms. In *2009 Information Theory and Applications Workshop*, 96–104.

Feng, S., Zhang, Y., Li, S.E., Cao, Z., Liu, H.X., and Li, L. (2019). String stability for vehicular platoon control: Definitions and analysis methods. *Annual Reviews in Control*, 47, 81–97.

Pates, R. (2015). A loopshaping approach to controller design in networks of linear systems. In *54th IEEE Conference on Decision and Control*, 6276–6281.

Pates, R. and Yamamoto, K. (2018). Scale free bounds on the amplification of disturbances in mass chains. In *2018 Annual American Control Conference (ACC)*, 6002–6005.

Seiler, P., Pant, A., and Hedrick, K. (2004). Disturbance propagation in vehicle strings. *IEEE Transactions on Automatic Control*, 49(10), 1835–1842.

Usmani, R.A. (1994). Inversion of a tridiagonal Jacobi matrix. *Linear Algebra and its Applications*, 212(213), 413–414.

Yamamoto, K. and Smith, M.C. (2016). Bounded disturbance amplification for mass chains with passive interconnection. *IEEE Transactions on Automatic Control*, 61(6), 1565–1574.

Yamamoto, K. (2021). Scale-invariant controller synthesis in mass chains. *IFAC-PapersOnLine*, 54(9), 78–83.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Exploring the Limits of Open Quantum Dynamics II: Gibbs-Preserving Maps from the Perspective of Majorization ⋆

Frederik vom Ende *

* *Dept. Chem., Lichtenbergstraße 4, 85747 Garching, Germany & Munich Centre for Quantum Science and Technology (MCQST), Schellingstraße 4, 80799 München, Germany (e-mail: frederik.vom-ende@tum.de).*

date: 14 June 2022

Abstract: Motivated by reachability questions in coherently controlled open quantum systems coupled to a thermal bath, as well as recent progress in the field of thermo-/vector-majorization we generalize classical majorization from unital quantum channels to channels with an arbitrary fixed point D of full rank. Such channels preserve some Gibbs-state and thus play an important role in the resource theory of quantum thermodynamics, in particular in thermo-majorization.

Based on this we investigate D-majorization on matrices in terms of its topological and order properties, such as existence of unique maximal and minimal elements, etc. Moreover we characterize D-majorization in the qubit case via the trace norm and elaborate on why this is a challenging task when going beyond two dimensions.

Keywords: Open quantum systems, quantum control theory, reachable sets, quantum thermodynamics, majorization

1. INTRODUCTION

Studying reachable sets of control systems is necessary to ensure well-posedness of a large class of (optimal) control tasks. In Dirr et al. (2019) toy models on the standard simplex of probability vectors were studied in order to answer reachability questions of controlled n-level systems coupled to a bath of finite temperature such that the coupling can be switched on and off. If the closed (unitary) part of the system can be fully controlled and the bath has temperature $T = 0$ then every quantum state [1] can be reached approximately from every initial state (that is, perhaps not exactly but at least with arbitrary precision). For $T = \infty$ an upper bound can be obtained by classical majorization techniques. For more details on this we refer to the first part of this talk: *Exploring the Limits of Open Quantum Dynamics I: Motivation, First Results from Toy Models to Applications*, as well as Section 3.3.

An obvious follow-up question is what can be said—if one can say anything at all—about the reachable set of such a system for $0 < T < \infty$? Even within the simplified diagonal toy model (cf. Part I) this is a rather difficult task and it seems that the notion necessary to handle such problems requires a more general form of majorization:

⋆ The project was supported i.a. by Excellence Network of Bavaria under ExQM and is part of *Munich Quantum Valley* of the Bavarian State Government with funds from Hightech Agenda *Bayern Plus*.

[1] A quantum state is a positive semi-definite matrix of unit trace.

2. ON THE ROAD TO D-MAJORIZATION

2.1 d-Majorization on Vectors

Majorization relative to a strictly positive vector $d \in \mathbb{R}^n_{++}$, as introduced by Veinott (1971) and in the quantum regime by Ruch et al. (1978) is defined as follows: a vector y is said to d-majorize x, denoted by $x \prec_d y$, if there exists a d-stochastic matrix $A \in \mathbb{R}^{n\times n}$ with $x = Ay$. Recall that $A \in \mathbb{R}^{n\times n}$ is d-stochastic if all its entries are non-negative and $Ad = d$, $e^\top A = e^\top$ with $e := (1, \dots, 1)^\top$. A variety of characterizations of $\prec_d$ and d-stochastic matrices can be found in the work of Joe (1990) or vom Ende and Dirr (2022). The most useful for numerical purposes is the following: $x \prec_d y$ if and only if $\sum_{j=1}^n x_j = \sum_{j=1}^n y_j$ and $\|d_i x - y_i d\|_1 \leq \|d_i y - y_i d\|_1$ for all $i = 1, \dots, n$, where $\|z\|_1 = \sum_{j=1}^n |z_j|$ is the usual vector-1-norm.

Classical majorization $\prec$, that is, $x \prec y$ for $x, y \in \mathbb{R}^n$, is originally defined via ordering x, y decreasingly and then comparing partial sums: $\sum_{j=1}^k x_j^{\cdot} \leq \sum_{j=1}^k y_j^{\cdot}$ for all $k = 1, \dots, n-1$ as well as $\sum_{j=1}^n x_j = \sum_{j=1}^n y_j$. For more on vector majorization we refer to Ch. 1 & 2 of Marshall et al. (2011). In particular it is well-known that setting $d = e$ in the definition of d-majorization recovers $\prec$ —which also shows that the definition via partial sums cannot extend beyond $e^\top$: as soon as two entries in $d \in \mathbb{R}^n_{++}$ differ one loses permutation invariance and reordering the vectors x, y makes a conceptual difference.

The above 1-norm characterization allows to rewrite the d-majorization polytope $M_d(y) := \{x \in \mathbb{R}^n \,|\, x \prec_d y\}$ for any

$y \in \mathbb{R}^n$ as the set of solutions to a nicely structured vector inequality $\mathcal{M}x \leq b$. Here $\mathcal{M} \in \mathbb{R}^{2^{n^2} \times n}$ is independent of y, d while the entries of $b = b(y,d) \in \mathbb{R}^{2^n}$ depend explicitly (and continuously) on y and d – for details cf. Thm. 10 in vom Ende and Dirr (2022). This description of d-majorization enables a proof of the existence of an extremal point $z \in M_d(y)$ such that $M_d(y) \subseteq M_e(z)$, i.e. there exists some $z \prec_d y$ which classically majorizes all $x \in M_d(y)$. Due to this result d-majorization is suited to analyse reachable sets in the toy model (cf. Part I of this talk)—yet as soon as one considers n-level quantum systems one needs a similar concept on (density) matrices.

2.2 Generalizing d-Majorization to Matrices

Classical majorization on the level of hermitian matrices uses their "eigenvalue vector" $\boldsymbol{\lambda}(\cdot)$ arranged in any order with multiplicities counted. For $A, B \in \mathbb{C}^{n\times n}$ hermitian, A is said to be majorized by B if $\boldsymbol{\lambda}(A) \prec \boldsymbol{\lambda}(B)$, cf. Ando (1989). The most naïve approach to define D-majorization on matrices (with [2] $D = \mathrm{diag}(d)$ for some $d \in \mathbb{R}^n_{++}$) would be to replace $\prec$ by $\prec_d$ and leave the rest as it is. However such a definition is unfeasible because it depends on the arrangement of the eigenvalues in $\boldsymbol{\lambda}$, due to the lack of permutation invariance of d (unless $d = e$).

The most natural way out of this dilemma is to remember that classical majorization on matrices can be equivalently characterised via linear maps which are completely positive and trace-preserving (CPTP) and which have the identity matrix $\mathrm{id} = \mathrm{diag}(1, \ldots, 1)$ as a fixed point. Therefore it seems utmost reasonable to generalize d-majorization on square matrices as follows:

Definition 1. Given $n \in \mathbb{N}$ and $A, B \in \mathbb{C}^{n\times n}$ as well as a positive definite matrix $D \in \mathbb{C}^{n\times n}$ we say that A is D-majorized by B (denoted by $A \prec_D B$) if there exists a CPTP map Φ such that $\Phi(B) = A$ and $\Phi(D) = D$.

Such a definition is also justified by the following: given real vectors x, y and a positive vector $d \in \mathbb{R}^n_{++}$ one can show that $\mathrm{diag}(x) \prec_{\mathrm{diag}(d)} \mathrm{diag}(y)$ if and only if $x \prec_d y$. In other words the diagonal case reduces to d-majorization on vectors as expected.

Be aware that one *could* define matrix D-majorization via positive (instead of completely positive) trace-preserving maps, and that this would make a conceptual difference – unless $D \neq \mathrm{id}$ (Ando, 1989, Thm. 7.1), more on this at the end of Section 3.1. However, we defined D-majorization via CPTP maps because this class has a richer theory behind it and because it is the more natural choice if one comes from quantum information and control.

3. PROPERTIES OF D-MAJORIZATION

Using CPTP maps in Definition 1 also allows for a physical interpretation of D-majorization: Given some n-level system (with Hamiltonian $H_0 \in \mathbb{C}^{n\times n}$) coupled to a bath of some temperature $T > 0$, the Gibbs state (that is, the thermodynamic equilibrium state) of the system is given by

$$\rho^{H_0,T}_{\mathrm{Gibbs}} := \frac{\exp(-H_0/T)}{\mathrm{tr}(\exp(-H_0/T))} > 0\,.$$

Because every positive definite $n \times n$ matrix of unit trace is the Gibbs state of *some* n-level system this links D-majorization to Gibbs-preserving CPTP maps. Moreover in the high-temperature limit the above definition reduces to $\lim_{T\to\infty} \rho^{H_0,T}_{\mathrm{Gibbs}} = \frac{1}{n}\,\mathrm{diag}(1, \ldots, 1)$, which connects classical majorization to baths of infinite temperature.

3.1 Characterizations of D-Majorization

An important observation is that for any $A, B \in \mathbb{C}^{n\times n}$ and $D > 0$ one has $A \prec_D B$ if and only if $UAU^* \prec_{UDU^*} UBU^*$ for all unitary matrices $U \in \mathbb{C}^{n\times n}$. Thus we can w.l.o.g. assume that D is diagonal in the standard basis.

Now if one deals with qubits, i.e. two-dimensional systems, then D-majorization can be characterized as follows.

Proposition 2. Let $d \in \mathbb{R}^2_{++}$, $D = \mathrm{diag}(d)$ and $A, B \in \mathbb{C}^{2\times 2}$ hermitian be given. The following are equivalent.

(i) $A \prec_D B$

(ii) *There exists a positive trace-preserving map Φ with $\Phi(D) = D$ and $\Phi(B) = A$.*

(iii) $\|A - tD\|_1 \leq \|B - tD\|_1$ *for all* $t \in \mathbb{R}$ *with* $\|\cdot\|_1 = \mathrm{tr}\sqrt{(\cdot)^*(\cdot)}$ *being the trace norm.*

(iv) $\mathrm{tr}(A) = \mathrm{tr}(B)$ *and* $\|A - b_i D\|_1 \leq \|B - b_i D\|_1$ *for* $i = 1, 2$ *as well as for the generalized fidelity*

$$\left\|\sqrt{A - b_1 D}\sqrt{b_2 D - A}\right\|_1 \geq \left\|\sqrt{B - b_1 D}\sqrt{b_2 D - B}\right\|_1$$

Here $\sigma(D^{-1/2} B D^{-1/2}) = \{b_1, b_2\}$ ($b_1 \leq b_2$) with $\sigma(\cdot)$ being the spectrum.

Of course property (iv) is the closest to the 1-norm characterization of $\prec_d$ from Sec. 2.1 and, moreover, the key to easily check (e.g., on a computer) if some hermitian matrix D-majorizes another. Unfortunately *none* of these characterizations generalize to dimensions larger than 2 because the counterexample to the Alberti-Uhlmann theorem in higher dimensions, given by Heinosaari et al. (2012), pertains to our problem: Consider the hermitian matrices

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & -i \\ 0 & i & 2 \end{pmatrix} \quad B = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & i \\ 0 & -i & 2 \end{pmatrix} \quad D = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}. \quad (1)$$

Then $\sigma(D) = \{2, 2+\sqrt{2}, 2-\sqrt{2}\}$ so $D > 0$. Obviously, $B^\top = A$ and $D^\top = D$ so because the transposition map is well-known to be linear, positivity- and trace-preserving one has $\|A - tD\|_1 = \|(B - tD)^\top\|_1 = \|B - tD\|_1$ for all $t \in \mathbb{R}$. But there exists no CPTP map, i.e. no $\Phi \in Q(n)$ such that $\Phi(B) = A$ and $\Phi(D) = D$ as shown in (Heinosaari et al., 2012, Proposition 6). For now finding simple-to-verify conditions for $\prec_D$ beyond two dimensions remains an open problem.

3.2 Order Properties of D-Majorization

As is readily verified $\prec_d$ is a preorder but it is not a partial order—the same holds for $\prec_D$ and the counterexample which shows that $\prec_d$ is not a partial order transfers to the matrix case. Moreover, now, one can characterize minimal and maximal elements in this preorder.

[2] Here and henceforth $\mathrm{diag}(x) \in \mathbb{C}^{n\times n}$ is the matrix which has $x \in \mathbb{C}^n$ on its diagonal and the remaining entries are 0.

Theorem 3. Let $d \in \mathbb{R}^n_{++}$ be given and let

$$\mathfrak{h}_d := \{X \in \mathbb{C}^{n\times n} \,|\, X \text{ hermitian and } \operatorname{tr}(X) = e^\top d\}$$
$$\mathfrak{h}_d^+ := \{X \in \mathbb{C}^{n\times n} \,|\, X \geq 0 \text{ and } \operatorname{tr}(X) = e^\top d\}$$

be the trace hyperplane induced by d within the hermitian and the positive semi-definite matrices, respectively. The following statements hold.

(i) $D = \operatorname{diag}(d)$ is the unique minimal element in $\mathfrak{h}_d$ with respect to $\prec_D$.

(ii) $(e^\top d)e_k e_k^\top$ is maximal in $\mathfrak{h}_d^+$ with respect to $\prec_D$ where k is chosen such that d_k is minimal in d. It is the unique maximal element in $\mathfrak{h}_d^+$ with respect to $\prec_D$ if and only if d_k is the unique minimal element of d.

From a physical point of view this is precisely what one expects: from the state with the largest energy one can generate every other state (in an equilibrium-preserving manner) and there is no other state with this property.

3.3 Reachable Sets & D-Majorization

Let us finally connect our notion of D-majorization to the reachability questions we touched upon in the introduction. Markovian quantum control systems are generally modelled via a controlled GKSL-equation [Gorini et al. (1976); Lindblad (1976)]:

$$\dot\rho(t) = -i\Big[H_0 + \sum\nolimits_{j=1}^m u_j(t)H_j, \rho(t)\Big] - \gamma(t)\Gamma(\rho(t)) \quad (2)$$

with initial state $\rho(0) = \rho_0 \in \mathbb{C}^{n\times n}$, control Hamiltonians $H_1, \dots, H_m$, and control amplitudes $u_1, \dots, u_m, \gamma$. Here $\Gamma(\rho) := \sum_{j\in I}(\frac{1}{2}(V_j^*V_j\rho + \rho V_j^*V_j) - V_j\rho V_j^*)$ describes the dissipative effect on the system by means of the matrices $(V_j)_{j\in I} \subset \mathbb{C}^{n\times n}$ which in principle can be arbitrary.

Now given any n-level system described by a hermitian matrix $H_S \in \mathbb{C}^{n\times n}$ with spectral decomposition $\sum_{j=1}^n E_j|g_j\rangle\langle g_j|$, $E_1 \leq \dots \leq E_n$ and a bath of some temperature $T > 0$ the coupling of the system to said bath can be modelled by (2) if the generators of the dissipation $(V_j)_{j\in I}$ are chosen to be the modified ladder operators

$$\begin{aligned}\sigma_+^d &=: \sum\nolimits_{j=1}^{n-1} \sqrt{\frac{j(n-j)e^{-E_j/T}}{e^{-E_j/T} + e^{-E_{j+1}/T}}}\, |g_j\rangle\langle g_{j+1}| \\ \sigma_-^d &=: \sum\nolimits_{j=1}^{n-1} \sqrt{\frac{j(n-j)e^{-E_{j+1}/T}}{e^{-E_j/T} + e^{-E_{j+1}/T}}}\, |g_{j+1}\rangle\langle g_j|\,.\end{aligned} \quad (3)$$

In order to analyze the reachable set of (2) with $H_0 = H_S$ and dissipation generators σ_+^d, σ_-^d we (as in Section 2.1) define the set of all matrices which are D-majorized by some state ρ or a collection of states $S \subseteq \mathbb{C}^{n\times n}$:

$$\begin{aligned}M_D : \mathcal{P}(\mathbb{C}^{n\times n}) &\to \mathcal{P}(\mathbb{C}^{n\times n}) \\ S &\mapsto \bigcup\nolimits_{\rho\in S}\{X \in \mathbb{C}^{n\times n} \,|\, X \prec_D \rho\}\end{aligned}$$

with $\mathcal{P}$ being the power set and $M_D(X) := M_D(\{X\})$ for all $X \in \mathbb{C}^{n\times n}$. This operator is used to upper bound the reachable set of the "toy model" Λ_d (cf. Part I)[3] and is expected to do so in the matrix case, as well. Important properties of M_D are:

(i) $M_D(X)$ is convex for all $X \in \mathbb{C}^{n\times n}$.

(ii) If $P \subset \mathbb{C}^{n\times n}$ is compact, then $M_D(P)$ is compact.

(iii) If P is a collection of quantum states then $M_D(P)$ is star-shaped with respect to the Gibbs state $\frac{D}{\operatorname{tr}(D)}$.

(iv) When restricting M_D to the compact subsets of $\mathbb{C}^{n\times n}$ then M_D is non-expansive (so in particular continuous) with respect to the Hausdorff metric.

The last property formulates that for a system in the state ρ which is coupled to a bath of temperature $T \geq 0$, "small" changes in ρ cannot change the set of D-majorized states "too much".

Coming back to footnote 3, the crucial step in the proof is to identify an extreme point of $M_d(x_0)$ which is maximal w.r.t. classical majorization. While an extreme point analysis of the set of matrices $M_D(\rho_0)$ is way more difficult—as the convex polytope techniques from the vector case break down—the idea of a maximal extreme point might be equally useful in analyzing general open quantum control problems in the future.

4. CONNECTION TO THERMO-MAJORIZATION

Over the last few years, sparked by Brandão et al. (2015); Horodecki and Oppenheim (2013) and others [Gour et al. (2015); Lostaglio et al. (2018); Sagawa et al. (2021)] thermo-majorization has been a widely discussed and studied topic in quantum physics and in particular quantum thermodynamics. In the abelian case thermo-majorization, on a mathematical level, is described by vector d-majorization which begs the question of how to define thermo-majorization for general quantum states.

Indeed Faist et al. (2015) have shown that it makes a conceptual difference whether one defines thermo-majorization on non-diagonal states via Gibbs-preserving maps (i.e. CPTP maps having the Gibbs state $D > 0$ as a fixed point, cf. Definition 1) or if one restricts to the smaller class of thermal operations. The latter, given some Hamiltonian of the system H_S and a fixed bath temperature $T \geq 0$, are defined as follows, cf. also Lostaglio (2019):

Definition 4. A linear map $\Phi : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is a thermal operation w.r.t. H_S if there exist $m \in \mathbb{N}$, $H_R \in \mathbb{C}^{m\times m}$ hermitian, and $U \in \mathbb{C}^{mn\times mn}$ unitary such that

$$[U, H_S \otimes \mathrm{id}_R + \mathrm{id}_S \otimes H_R] = 0 \quad (4)$$

and

$$\Phi(\rho) = \operatorname{tr}_R(U(\rho \otimes \rho_{\text{Gibbs}}^{H_R,T})U^*)$$

for all $\rho \in \mathbb{C}^{n\times n}$ (or equivalently for all quantum states ρ). We denote the collection of all thermal operations by $\mathsf{TO}(H_S, T)$.

Thermal operations are the free operations of the resource theory of quantum thermodynamics as those encompass the dynamics which preserve the Gibbs-state and which satisfy (4) (conserve the global energy, that is, the energy of the larger system $H_{SR} = H_S \otimes \mathrm{id}_R + \mathrm{id}_S \otimes H_R$). One readily verifies that $\mathsf{TO}(H_S, T)$ forms a path-connected semigroup with identity and – although $\mathsf{TO}(H_S, T)$ in general is not closed – its closure even is convex and compact.

In the vector case the state transitions possible with thermal operations are the same as with general Gibbs-

[3] More precisely we proved that $\mathfrak{reach}_{\Lambda_d}(x_0) \subseteq (M_e \circ M_d)(x_0)$ for any initial state x_0 and $d \in \mathbb{R}^n_{++}$ corresponding to a spin system, i.e. $d = (\alpha^{j-1})_{j=1}^n$ for some $\alpha \in (0, 1)$.

preserving maps described by d-stochastic matrices. However in the operator case there is a discrepancy between the two coming from the fact that there exist Gibbs-preserving maps which generate coherent superpositions of energy levels, whereas no thermal operation is capable of doing such a thing. In fact for all $H_S \in \mathbb{C}^{n\times n}$, $T \geq 0$ one finds the inclusions

$$\mathsf{TO}(H_S, T) \subseteq \mathsf{EnTO}(H_S, T) \subsetneq Q_{e^{-H_S/T}}(n) \qquad (5)$$

where $Q_{e^{-H_S/T}}(n)$ is the collection of all CPTP maps which have $e^{-H_S/T}$ – and thus $\rho^{H_S,T}_{\text{Gibbs}}$ – as a fixed point, and

$$\mathsf{EnTO}(H_S, T) := \{\Phi \in Q_{e^{-H_S/T}}(n) \,:\, [\Phi, \mathrm{ad}_{H_S}] = 0\}$$

are the enhanced thermal operations (also called "covariant Gibbs-preserving maps"). This is an important observation as the covariance property $[\Phi, \mathrm{ad}_{H_S}] = 0$ forces that the diagonal and the off-diagonal action of any channel are strictly separated, assuming H_S has non-degenerate spectrum. Note that this insight is of importance to us because the solution to the uncontrolled master equation (2) (i.e. $H_1 = \ldots = H_m = 0$, $\gamma \equiv 1$) with dissipation generators σ^d_+, σ^d_- from (3) lives in EnTO at all times.

Be aware that in (5), if the thermal operations are replaced by their closure then the first set inclusion is an equality if $n = 2$ and becomes a strict inclusion for $n \geq 3$ as shown by Ding et al. (2021). Even worse this discrepancy between TO and EnTO remains when looking at the *action* of the respective sets on certain states; more precisely, there exist quantum states $\rho, \omega \in \mathbb{C}^{3\times 3}$ and an enhanced thermal operation Φ such that $\Phi(\rho) = \omega$ but no element in TO or its closure can map ρ to ω.

This observation is particularly important for the field of quantum control as there one usually wonders which state transitions can be realized under a given control scenario. Thus beyond qubits it makes a conceptional difference which of the sets in (5) one uses to model a given thermodynamic control problem. Moreover, quantum control problems usually come in the framework of quantum-dynamical semigroups so one in addition needs to identify those quantum maps from a certain set (usually carrying the structure of a semigroup) which can be written as the solution to a controlled master equation of Gorini-Kossakowski-Sudarshan-Lindblad type [Gorini et al. (1976); Lindblad (1976)], that is, to identify those channels which are time-dependent Markovian. In other words from a Lie-theoretical perspective one wants to determine the Lie wedge of the respective semigroup in order to characterize the desired quantum channels as solutions of suitable (bi)linear master equations.

The question of Markovian state transitions in thermodynamics has only been tackled recently by Lostaglio and Korzekwa (2021) for the set of enhanced thermal operations and the simpler case of diagonal states – recall that in this realm the problem reduces to vector-d majorization. They were able to fully characterize which state transitions are possible under Markovian thermal processes (i.e. maps from EnTO which are solutions of a time-dependent GKSL-equation) in the classical realm, and they even gave algorithms to check for a Markovian path from a given initial to a given final state. While incredibly important, their work of course is but a first step in this direction and the ultimate goal will be to extend their results and concepts to general thermodynamic quantum control systems.

REFERENCES

Ando, T. (1989). Majorization, Doubly Stochastic Matrices, and Comparison of Eigenvalues. *Lin. Alg. Appl.*, 118, 163–248.

Brandão, F., Horodecki, M., Ng, N., Oppenheim, J., and Wehner, S. (2015). The Second Laws of Quantum Thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 112, 3275–3279.

Ding, Y., Ding, F., and Hu, X. (2021). Exploring the Gap Between Thermal Operations and Enhanced Thermal Operations. *Phys. Rev. A*, 103, 052214.

Dirr, G., vom Ende, F., and Schulte-Herbrüggen, T. (2019). Reachable Sets from Toy Models to Controlled Markovian Quantum Systems. *Proc. IEEE Conf. Decision Control (IEEE-CDC)*, 58, 2322.

Faist, P., Oppenheim, J., and Renner, R. (2015). Gibbs–Preserving Maps Outperform Thermal Operations in the Quantum Regime. *New J. Phys.*, 17, 1–4.

Gorini, V., Kossakowski, A., and Sudarshan, E. (1976). Completely Positive Dynamical Semigroups of N-Level Systems. *J. Math. Phys.*, 17, 821–825.

Gour, G., Müller, M., Narasimhachar, V., Spekkens, R., and Halpern, N. (2015). The Resource Theory of Informational Nonequilibrium in Thermodynamics. *Phys. Rep.*, 583, 1–58.

Heinosaari, T., Jivulescu, M., Reeb, D., and Wolf, M. (2012). Extending Quantum Operations. *J. Math. Phys.*, 53, 102208.

Horodecki, M. and Oppenheim, J. (2013). Fundamental Limitations for Quantum and Nanoscale Thermodynamics. *Nat. Commun.*, 4, 2059.

Joe, H. (1990). Majorization and Divergence. *J. Math. Anal. Appl.*, 148, 287–305.

Lindblad, G. (1976). On the Generators of Quantum Dynamical Semigroups. *Commun. Math. Phys.*, 48, 119–130.

Lostaglio, M. (2019). An Introductory Review of the Resource Theory Approach to Thermodynamics. *Rep. Prog. Phys.*, 82, 114001.

Lostaglio, M., Alhambra, Á., and Perry, C. (2018). Elementary Thermal Operations. *Quantum*, 2, 1–52.

Lostaglio, M. and Korzekwa, K. (2021). Continuous Thermomajorization and a Complete Set of Laws for Markovian Thermal Processes.

Marshall, A., Olkin, I., and Arnold, B. (2011). *Inequalities: Theory of Majorization and Its Applications*. Springer, New York, 2 edition.

Ruch, E., Schranner, R., and Seligman, T. (1978). The Mixing Distance. *J. Chem. Phys.*, 69, 386–392.

Sagawa, T., Faist, P., Kato, K., Matsumoto, K., Nagaoka, H., and Brandão, F. (2021). Asymptotic Reversibility of Thermal Operations for Interacting Quantum Spin Systems via Generalized Quantum Stein's Lemma. *J. Phys. A*, 54, 495303.

Veinott, A. (1971). Least d-Majorized Network Flows with Inventory and Statistical Applications. *Manag. Sci.*, 17, 547–567.

vom Ende, F. and Dirr, G. (2022). The d-Majorization Polytope. *Lin. Alg. Appl.*, 649, 152–185.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Exploring the Limits of Open Quantum Dynamics I: Motivation, New Results from Toy Models to Applications ⋆

Thomas Schulte-Herbrüggen * **Frederik vom Ende** *
Gunther Dirr **

* *Dept. Chem., Lichtenbergstraße 4, 85747 Garching, Germany & Munich Centre for Quantum Science and Technology (MCQST), (e-mail: {tosh, frederik.vom-ende}@tum.de).*
** *Mathematics Inst., University of Würzburg, Emil-Fischer-Straße 40, 97074 Würzburg, Germany, (e-mail: dirr@mathematik.uni-wuerzburg.de)*

Abstract: Which quantum states can be reached by controlling open Markovian n-level quantum systems? Here, we address reachable sets of coherently controllable quantum systems with switchable coupling to a thermal bath of temperature T. — The core problem reduces to a toy model of studying points in the standard simplex allowing for two types of controls: (i) permutations within the simplex, (ii) contractions by a dissipative semigroup [Dirr et al. (2019)]. By illustration, we put the problem into context and show how toy-model solutions pertain to the reachable set of the original controlled Markovian quantum system. Beyond the case $T = 0$ (amplitude damping) we present new results for $0 < T < \infty$ using methods of d-majorisation.

Keywords: Quantum Control Theory; Markovian Quantum Dynamics; Reachable Sets; Quantum Thermodynamics; Majorisation, d-Majorisation.

1. INTRODUCTION

Here we show how reachability problems of (finite dimensional) Markovian open quantum systems may reduce to hybrid control systems on the standard simplex of $\mathbb{R}^n$. Consider a bilinear control system [Jurdjevic (1997); Elliott (2009)]

$$\dot{\mathbf{x}}(t) = -(A + \sum\nolimits_j u_j(t) B_j)\mathbf{x}(t)\,, \quad \mathbf{x}(0) = \mathbf{x}_0\,, \qquad (1)$$

where as usual A denotes an uncontrolled drift, while the control terms consist of (piecewise constant) control amplitudes $u_j(t) \in \mathbb{R}$ and control operators B_j. The state $\mathbf{x}(t)$ may be thought of as (vectorized) density operator. The corresponding system Lie algebra, which provides the crucial tool for analysing controllability and accessibility questions, reads $\mathfrak{k} := \langle A, B_j \,|\, j = 0, 1, \ldots, m\rangle_{\mathsf{Lie}}$.

For 'closed' quantum systems, i.e. systems which do not interact with their environment, the matrices A and B_j involved are skew-hermitian and thus it is known [Jurdjevic and Sussmann (1972); Brockett (1972); Jurdjevic (1997)] that the reachable set of (1) is given by the orbit of the initial state under the action of the dynamical systems group $\mathbf{K} := \langle \exp \mathfrak{k} \rangle$, provided $\mathbf{K}$ is a closed and thus compact subgroup of the unitary group.

More generally, for 'open' systems undergoing Markovian dissipation, the reachable set takes the form of a (Lie) semigroup orbit, see, e.g., [Dirr et al. (2009)]. – Here we address a scenario with coherent controls $\{B_j\}_{j=1}^m$ and a *bang-bang switchable* dissipator B_0 as motivated by recent experimental progress [Chen et al. (2014); Wong et al. (2019)] and described in Bergholm et al. (2016).

Specification of the Toy Model — These assumptions and the 'thermal' condition (*vide infra*) that B_0 leaves the set of diagonal matrices invariant simplify the reachability analysis of (1) to the core problem of diagonal states represented by probability vectors of the standard simplex

$$\Delta^{n-1} := \left\{ x \in \mathbb{R}^n_+ \,|\, \textstyle\sum_{i=1}^n x_i = 1 \right\}, \qquad (2)$$

i.e. $\mathbf{x}(t) = \mathrm{diag}(x(t))$. In order to make the main features match the quantum dynamical context, let us fix the following stipulations for the toy model: Its controls shall amount to permutation matrices acting instantaneously on the entries of $x(t)$ and a continuous-time one-parameter semigroup $(e^{-tB_0})_{t\in\mathbb{R}^+}$ of stochastic maps with a unique fixed point d in Δ^{n-1}. As $(e^{-tB_0})_{t\in\mathbb{R}^+}$ results from the restriction of the bang-bang switchable dissipator B_0, with abuse of notation we will denote its infinitesimal generator again by B_0. The '*equilibrium state*' d is defined in (8) by system parameters and the absolute temperature $T \geq 0$ of an external bath.

These stipulations suggest the following hybrid/impulsive scenario to define the 'toy model' Λ on $\Delta^{n-1} \subset \mathbb{R}^n$ by

$$\begin{aligned} &\dot{x}(t) = -B_0 x(t)\,, \quad x(t_k) = \pi_k x_k\,, \quad t \in [t_k, t_{k+1})\,, \\ &x_0 \in \Delta^{n-1}\,, \quad x_{k+1} = e^{-(t_{k+1}-t_k)B_0} x(t_k)\,, \quad k \geq 0\,. \end{aligned} \qquad (3)$$

Furthermore, $0 =: t_0 \leq t_1 \leq t_2 \leq \ldots$ is an arbitrary switching sequence and π_k are arbitrary permutation ma-

⋆ The project was supported i.a. by Excellence Network of Bavaria under ExQM and is part of *Munich Quantum Valley* of the Bavarian State Government with funds from Hightech Agenda *Bayern Plus*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

trices. Both the switching points and the permutation matrices are regarded as controls for (3). For simplicity, we assume that the switching points do not accumulate on finite intervals. For more details on hybrid/impulsive control systems see, e.g., [Lakshmikantham et al. (1989); Alur et al. (1996)]. The reachable sets of (3)

$$\mathfrak{reach}_\Lambda(x_0) := \{x(t) \,|\, x(\cdot) \text{ is a solution of (3), } t \geq 0\}$$

allow for the characterisation $\mathfrak{reach}_\Lambda(x_0) = \mathcal{S}_\Lambda x_0$, where $\mathcal{S}_\Lambda \subset \mathbf{L}(n,\mathbb{R})$ is the contraction semigroup generated by $(e^{-tB_0})_{t\in\mathbb{R}_+}$ and the set of all permutation matrices π.

2. STATE-OF-THE-ART

Henceforth, let $\mathbf{\Gamma}$ stand for a GKSL-operator acting on complex $n \times n$ matrices, see (5). Then A in (1) can be regarded as its matrix representation (e.g., in the Kronecker formalism (Horn and Johnson, 1991, Chap. 4)). If $\mathbf{\Gamma}$ leaves the set of diagonal matrices invariant—as in *enhanced thermal operations* [Lostaglio (2019)] we use as controllable resource—its reduction to the action on diagonal states represented as vectors is denoted B_0 since it originates from *switchable* noise. — In this picture, our recent results [Dirr et al. (2019)] can be sketched as follows.

Take the n-level toy model $\Lambda := \Lambda_0$ with controls by permutations and an infinitesimal generator B_0 which results from coupling to a bath of temperature $T = 0$ (i.e. $\mathbf{\Gamma} := \mathbf{\Gamma}_0$ given by single $V := \sigma_-$ see (10) with $\theta = \frac{\pi}{2}$).

Theorem 1. The closure of the reachable set of any initial vector $x_0 \in \Delta^{n-1}$ under the dynamics of Λ_0 exhausts the full standard simplex, i.e. $\overline{\mathfrak{reach}_{\Lambda_0}(x_0)} = \Delta^{n-1}$.

Moving from a single n-level system (qu*d*it) with $x_0 \in \Delta^{n-1}$ to a tensor product of m such n-level systems gives $x_0 \in \Delta^{n^m-1} \subset (\mathbb{R}^n)^{\otimes m}$. If the bath of temperature $T = 0$ is coupled to just one (say the last) of the m qu*d*its, $\mathbf{\Gamma}_0$ is generated by $V := I_{n^{m-1}} \otimes \sigma_-$ and one obtains the following generalization.

Theorem 2. The statement of Theorem 1 holds analogously for all m-qudit states $x_0 \in \Delta^{n^m-1}$.

In a first round to generalise the findings from the extreme cases $T = 0$ or $T = \infty$ to finite temperatures $0 < T < \infty$ we found the following: Let $\mathbf{\Gamma} := \mathbf{\Gamma}_d$ be the dissipator for temperature $T > 0$ with $\mathbf{\Gamma}_d$ comprising the generators σ_-^d and σ_+^d of (9) and (10) as detailed in Sec. 4 and let $d \in \Delta^{n-1}$ be its unique attractive fixed point given by (8). For equidistant eigenvalues in H_0 (see Sec. 5) one gets:

Theorem 3. Again allowing for permutations as controls interleaved with dissipation resulting from $B_0(\mathbf{\Gamma}_d)$ one obtains for the reachable set of the thermal state d under the dynamics of the respective toy model $\Lambda := \Lambda_d$ the inclusion $\mathfrak{reach}_{\Lambda_d}(d) \subseteq \{x \in \Delta^{n-1} \,|\, x \prec d\}$, where '$\prec$' refers to the standard concept and notation of majorisation [Marshall et al. (2011); Ando (1989)].

Our recent toy-model results in Dirr et al. (2019) thus extend (the diagonal part of) the qu*b*it picture previously analysed by Bergholm et al. (2016) to n-level systems, and even more generally to systems of m qu*d*its. Here we explore further generalisations to finite temperatures $0 < T < \infty$, e.g., by allowing for general initial states x_0 instead of the thermal state d in Theorem 3.

3. RELATION TO CONTROLLED QUANTUM MARKOVIAN DYNAMICS

Let $\mathcal{D}(n)$ denote all $n \times n$ density matrices (positive semi-definite with trace 1) and $\mathcal{L}(\mathbb{C}^{n\times n})$ be the space of all linear operators acting on complex $n \times n$-matrices. Then

$$\dot\rho(t) = -\mathbf{\Gamma}(\rho(t)), \quad \rho(0) = \rho_0 \in \mathcal{D}(n) \tag{4}$$

with $\mathbf{\Gamma} \in \mathcal{L}(\mathbb{C}^{n\times n})$ of the GKSL-form [Gorini et al. (1976); Lindblad (1976)] with $V_k \in \mathbb{C}^{n\times n}$ chosen arbitrary in

$$\mathbf{\Gamma}(\rho) := \sum\nolimits_k \left(\tfrac{1}{2}\big(V_k^\dagger V_k \rho + \rho V_k^\dagger V_k\big) - V_k \rho V_k^\dagger\right) \tag{5}$$

ensures the time evolution $\rho(t) = e^{-t\mathbf{\Gamma}}\rho_0$ solving (4) remains in $\mathcal{D}(n)$ for all $t \in \mathbb{R}_+$. So $(e^{-t\mathbf{\Gamma}})_{t\in\mathbb{R}_+}$ is a completely positive trace-preserving (i.e. CPTP) linear contraction semigroup leaving $\mathcal{D}(n)$ invariant.

The overarching goal is to characterise control systems Σ extending (4) by coherent controls (generated by hermitian H_j and (piece-wise constant) $u_j(t) \in \mathbb{R}$) and by making dissipation bang-bang switchable in the sense

$$\dot\rho(t) = -\mathrm{i}\Big[H_0 + \sum_{j=1}^m u_j(t) H_j, \rho(t)\Big] - \gamma(t)\mathbf{\Gamma}(\rho(t)) \tag{6}$$

with $\gamma(t) \in \{0, 1\}$. A general analytic description of reachable sets of (6) is challenging in particular in higher dimensional cases except for a few scenarios which allow explicit characterizations: (a) In the unital case $\mathbf{\Gamma}(I_n) = 0$, one has [Ando (1989); Yuan (2010)]

$$\mathfrak{reach}_\Sigma(\rho_0) \subseteq \{\rho \in \mathcal{D}(n) \,|\, \rho \prec \rho_0\}. \tag{7}$$

(b) If in addition $\mathbf{\Gamma}$ is generated by a single normal V, one gets (up to closure) equality in (7) provided the unitary part of (6) is *controllable* and the switching function $\gamma(t)$ gives extra control (cf. [Bergholm et al. (2016)] for finite and [vom Ende et al. (2019)] for infinite dimensions).

Under the controllability scenario given in (b) plus the invariance of diagonal states imposed by enhanced thermal operations [Lostaglio (2019)], the closure of the unitary orbit of $\mathrm{diag}\big(\mathfrak{reach}_\Lambda(x_0)\big)$ sits in the closure of $\mathfrak{reach}_\Sigma(U\,\mathrm{diag}(x_0)U^\dagger)$. Settings beyond thermal relaxation are pursued with similar techniques, e.g., by Rooney et al. (2018) at the expense of arriving at conditions that are hard to verify for higher-dimensional systems.

4. THERMAL STATES AND d-MAJORISATION

By unitary controllability choose H_0 diagonal (with increasing eigenvalues ϵ_k), so the equilibrium state resulting from coupling to a bath of temperature T is the *Gibbs vector*

$$d = \frac{(e^{-\epsilon_k/T})_{k=1}^n}{\sum_{k=1}^n e^{-\epsilon_k/T}} \in \Delta^{n-1} \tag{8}$$

with $\rho_{\mathsf{Gibbs}} = \mathrm{diag}(d) \in \mathcal{D}(n)$. As shown in Dirr et al. (2019), $\mathrm{diag}(d)$ can then be obtained as the unique fixed point of (4) when choosing the two Lindblad terms as

$$V_1 = \sigma_+^d := \sum\nolimits_{k=1}^{n-1} \sqrt{k(n-k)}\cos(\theta_k)\, E_{k,k+1} \tag{9}$$

$$V_2 = \sigma_-^d := \sum\nolimits_{k=1}^{n-1} \sqrt{k(n-k)}\sin(\theta_k)\, E_{k+1,k}, \tag{10}$$

where the $E_{i,j}$ denote standard Weyl matrices and

$$\theta_k := \arccos\sqrt{1 + d_{k+1}/d_k} \in (0, \tfrac{\pi}{2}). \tag{11}$$

Since diagonal states remain diagonal under enhanced thermal operations as with $\mathbf{\Gamma} := \mathbf{\Gamma}_d$ and V_1, V_2 above, the connection to the toy model Λ_d is obvious. — What one can do with thermal resources is determined via d-majorisation (a.k.a. thermomajorisation in Horodecki and Oppenheim (2013) or Brandão et al. (2015)) as generalisation of common majorisation [Marshall et al. (2011)] to majorisation with respect to a strictly positve vector d [Veinott (1971)] representing the 'Gibbs state':

Definition 1. For $x, y \in \mathbb{R}^n$ and $d \in \Delta^{n-1}$, the vector x is d-majorised by y, written $x \prec_d y$, if there is a column stochastic matrix $A \in \mathbb{R}^{n\times n}$ (all elements non-negative, columns summing up to one) with $Ad = d$ and $Ay = x$.

Note that d-majorisation reproduces conventional majorisation with A being doubly stochastic if d is the maximally mixed state $d = \frac{1}{n}e$ and e is the vector with all entries 1.

For numerics a convenient equivalent characterisation [vom Ende and Dirr (2022)] is: $x \prec_d y$ if and only if

$$(a)\ \ \Sigma_i x_i = \Sigma_i y_i \quad \text{and} \tag{12}$$

$$(b)\ \ \|d_i\, x - y_i\, d\|_1 \leq \|d_i\, y - y_i\, d\|_1 \ \forall\, i \in \{1, \ldots, n\}, \tag{13}$$

where $\|z\|_1 := \sum_{i=1}^n |z_i|$ is the vector 1-norm.

5. OVERVIEW OF NEW RESULTS

To motivate the study of the d-majorisation polytope (and its operator lift) in Part II, we illustrate new results for n-level systems by dynamics of three-level systems (qutrits) in the toy-model scenario with its population flips and its dynamics by coupling the system to a bath of various temperatures $0 \leq T \leq \infty$ entailing unique equilibrium states d given by (8). Henceforth we invoke

Assumption A: H_0 has equidistant energy eigenvalues.

Moreover define the set of vectors in the simplex Δ^{n-1} that are d-majorised by the initial state x_0 as

$$\Delta_d^{n-1}(x_0) := \{z \in \Delta^{n-1} \,|\, z \prec_d x_0\}\,, \tag{14}$$

while those conventionally majorised by x_0 shall be denoted as $\Delta_e^{n-1}(x_0)$. For the toy-model dynamics one gets:

(1) $e^{-tB_0}x_0 \in \Delta_d^{n-1}(x_0)$ for all $t \geq 0$;
(2) $\Delta_d^{n-1}(x_0)$ is a convex subset within the simplex Δ^{n-1},

which means the *dissipative time evolution* of any x_0 remains within the convex set of states d-majorised by x_0.

Beyond pure dissipative evolution the toy model also allows for permutations π, so one naturally obtains

$$\mathfrak{reach}_{\Lambda_d}(x_0) = \mathfrak{reach}_{\Lambda_d}(\pi(x_0)) \quad \forall \pi \in \mathcal{S}_n\,. \tag{15}$$

Clearly, the simplex region $\Delta_d^{n-1}(x_0)$ intertwines overall permutations π (in the symmetric group $\mathcal{S}_n$) in the sense

$$\pi\, \Delta_d^{n-1}(x_0) = \Delta_{\pi(d)}^{n-1}(\pi(x_0))\,. \tag{16}$$

For the maximally mixed state ($d \simeq e$) this boils down to permutation invariance under conventional majorisation

$$\pi\, \Delta_e^{n-1}(x_0) = \Delta_e^{n-1}(\pi(x_0)) = \Delta_e^{n-1}(x_0)\,. \tag{17}$$

Eq. (16) entails as a first new result:

Theorem 4. (generalising Thm. 3). Assuming **A** those initial states $\tilde{x}_0$ conventionally majorised by d (i.e. $\tilde{x}_0 \in \Delta_e^{n-1}(d)$) remain within $\Delta_e^{n-1}(d)$ under the dynamics of the toy model Λ_d. In other words $\mathfrak{reach}_{\Lambda_d}(\tilde{x}_0) \subseteq \Delta_e^{n-1}(d)$.

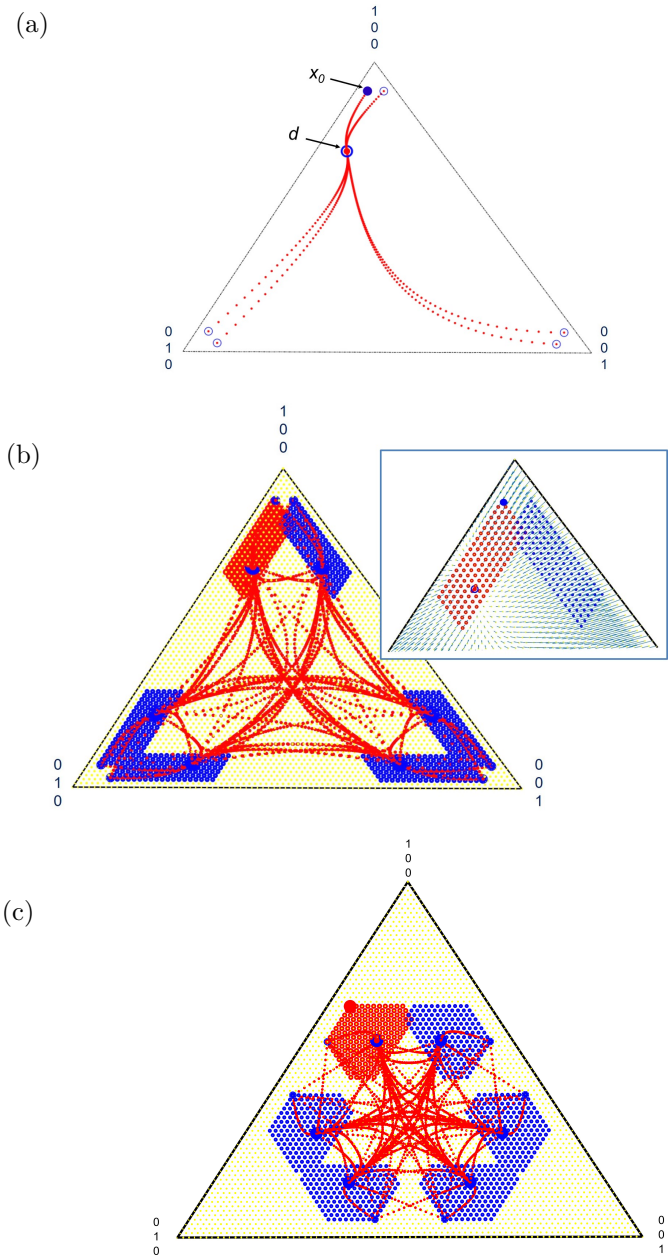


Fig. 1. **(a)** Evolutions of initial state $x_0 = (0.9, 0.07, 0.03)^\top$ and its permutations $\pi(x_0)$ under $\mathbf{\Gamma}_d$ with V_1, V_2 of (9) and $\theta = \frac{\pi}{6}$ of (11) drive into fixed point d;
(b) includes all permutations of trajectories starting with permutations of d, i.e. $x_0 = \pi(d)$; red region shows the states d-majorised by x_0, blue regions are their permutations; the convex hull over red and blue regions embraces entire reachable set $\mathfrak{reach}_{\Lambda_d}(x_0)$; inset gives the vector field to the dynamics Λ_d.
(c) For $\theta = \frac{\pi}{5}$ in (11) as in the generic case, the extreme point $z = (0.65, 0.30, 0.05)^\top$ in red differs from $x_0 = (0.55, 0.40, 0.05)^\top$ and $d = (0.55, 0.29, 0.16)^\top$.

In the next step, writing $x_0^\downarrow$ for ordering the entries of x_0 in descending magnitude (so that $x_0^\downarrow$ and d—with d being

the thermal state hence sorted by descending entries—are in the same Weyl chamber), one arrives at:

Theorem 5. Under assumption **A** the reachable set of the dynamics Λ_d is included in the set of all states conventionally majorised by $\Delta_d^{n-1}(x_0^{\downarrow})$ in the formal sense

$$\mathfrak{reach}_{\Lambda_d}(x_0) \subseteq \Delta_e^{n-1}\big(\Delta_d^{n-1}(x_0^{\downarrow})\big)\,. \tag{18}$$

The proof uses two facts: (i) There exists a (unique) extreme point z of the d-majorisation polytope $\Delta_d^{n-1}(x_0^{\downarrow})$ which conventionally majorises all points in $\Delta_d^{n-1}(x_0^{\downarrow})$, i.e. $\Delta_d^{n-1}(x_0^{\downarrow}) \subset \Delta_e^{n-1}(z)$. (ii) The vector field driving the dynamics of Λ_d points *inside* the conventional majorisation polytope $\Delta_e^{n-1}(z)$ at each of its $n!$ extreme points $\pi(z)$ with $\pi \in \mathcal{S}_n$ (cf. Fig. 1(c)). Once knowing how to construct z (see Part-II and [vom Ende and Dirr (2022)] for more detail), the results may be summarised and significantly simplified from d-majorisation to conventional majorisation via the extremal state z:

Theorem 6. Invoke assumption **A**. Then for the toy model Λ_d with Gibbs state d the reachable set is included in the following convex hull

$$\mathfrak{reach}_{\Lambda_d}(x_0) \subseteq \operatorname{conv}\big\{\pi(z)\,|\,\pi \in \mathcal{S}_n\big\} = \Delta_e^{n-1}(z)\,. \tag{19}$$

Fig. 1 illustrates these findings in three-level systems again assuming equidistant separation of energy eigenvalues for the underlying drift term H_0.

Conclusion and Outlook — For any initial state x_0, the time evolutions of probability vectors $x(t)$ following the underlying toy model Λ_d (thermal relaxation interdispersed with level-permutation) remain within the convex hull of extreme points resulting from the set of all states d-majorised by the initial state x_0. Yet, upon moving from the toy model to the full quantum dynamics of thermal relaxation interdispersed with unitary coherent evolution, the scenario gets more involved. In accompanying studies presented at this conference we pursue different approaches to handle the general case: (i) lifting d-majorisation to the operator level which leads to the concept of D-majorisation and (ii) projecting the full dynamics along the unitary orbits to the standard simplex by symmetric Lie algebra techniques giving an enhanced toy model for the reduced control system on Δ^{n-1}.

ACKNOWLEDGEMENTS

Fruitful discussion with David Reeb on unital systems at a very early phase of the project is gratefully acknowledged.

REFERENCES

Alur, R., Henzinger, T.A., and Sontag, E.D. (1996). *Hybrid Systems III: Verification and Control*. Lecture Notes in Computer Science, Vol. 1066. Springer, New York.

Ando, T. (1989). Majorization, Doubly Stochastic Matrices, and Comparison of Eigenvalues. *Lin. Alg. Appl.*, 118, 163.

Bergholm, V., Wilhelm, F., and Schulte-Herbrüggen, T. (2016). Arbitrary n-Qubit State Transfer Implemented by Coherent Control and Simplest Switchable Local Noise. `https://arxiv.org/abs/1605.06473v2`.

Brandão, F., Horodecki, M., Ng, N., Oppenheim, J., and Wehner, S. (2015). The Second Laws of Quantum Thermodynamics. *Proc.Natl.Acad.Sci. USA*, 112, 3275.

Brockett, R.W. (1972). System Theory on Group Manifolds and Coset Spaces. *SIAM J. Control*, 10, 265.

Chen, Y., et mult. al., and Martinis, J.M. (2014). Qubit Architecture with High Coherence and Fast Tunable Coupling. *Phys. Rev. Lett*, 113, 220502.

Dirr, G., Helmke, U., Kurniawan, I., and Schulte-Herbrüggen, T. (2009). Lie-Semigroup Structures for Reachability and Control of Open Quantum Systems: Kossakowski-Lindblad Generators form Lie Wedge to Markovian Channels. *Rep. Math. Phys.*, 64, 93.

Dirr, G., vom Ende, F., and Schulte-Herbrüggen, T. (2019). Reachable Sets from Toy Models to Controlled Markovian Quantum Systems. *Proc. IEEE Conf. Decision Control (IEEE-CDC)*, 58, 2322. `https://arxiv.org/abs/1905.01224`.

Elliott, D. (2009). *Bilinear Control Systems: Matrices in Action*. Springer, London.

Gorini, V., Kossakowski, A., and Sudarshan, E. (1976). Completely Positive Dynamical Semigroups of N-Level Systems. *J. Math. Phys.*, 17, 821.

Horn, R.A. and Johnson, C.R. (1991). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge.

Horodecki, M. and Oppenheim, J. (2013). Fundamental Limitations for Quantum and Nanoscale Thermodynamics. *Nat. Commun.*, 4, 2059.

Jurdjevic, V. (1997). *Geometric Control Theory*. Cambridge University Press, Cambridge.

Jurdjevic, V. and Sussmann, H. (1972). Control Systems on Lie Groups. *J. Diff. Equat.*, 12, 313.

Lakshmikantham, V., Bainov, D.D., and Simeonov, P.S. (1989). *Theory of Impulsive Differential Equations*. Series in Modern Applied Mathematics, Vol. 6. World Scientific, Singapore.

Lindblad, G. (1976). On the Generators of Quantum Dynamical Semigroups. *Commun. Math. Phys.*, 48, 119.

M. Lostaglio(2019). An Introductory Review of the Resource Theory Approach to Thermodynamics. *Rep. Prog. Phys.*, 82, 114001.

Marshall, A., Olkin, I., and Arnold, B. (2011). *Inequalities: Theory of Majorization and Its Applications*. Springer, New York, second edition.

Rooney, P., Bloch, A., and Rangan, C. (2018). Steering the Eigenvalues of the Density Operator in Hamiltonian-Controlled Quantum Lindblad Systems. *IEEE Trans. Automat. Control*, 63, 672.

Veinott, A. (1971). Least d-Majorized Network Flows with Inventory and Statistical Applications. *Manag. Sci.*, 17, 547.

vom Ende, F. and Dirr, G. (2022). The d-Majorization Polytope. *Lin. Alg. Appl.*, 649, 152.

vom Ende, F. , Dirr, G. , Keyl, M. and Schulte-Herbrüggen, T. (2019). Reachability in Infinite-Dimensional Unital Open Quantum Systems with Switchable GKS–Lindblad Generators. *Open Sys. Information Dyn.*, 26, 1950014.

Wong, C., Wilen, C., McDermott, R., and Vavilov, M. (2019). A Tunable Quantum Dissipator for Active Resonator Reset in Circuit QED. *Quant. Sci. Technol.*, 4, 025001.

Yuan, H. (2010). Characterization of Majorization Monotone Quantum Dynamics. *IEEE. Trans. Autom. Contr.*, 55, 955.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Extended differential balancing and generalized balancing for nonlinear dynamical systems

Arijit Sarkar * **Jacquelien M.A. Scherpen** *

* *Jan C Willems Center for Systems and Control, Discrete Technology and Production Automation, ENTEG, Faculty of Science and Engineering, University of Groningen (e-mail: {a.sarkar,j.m.a.scherpen}@rug.nl).*

Keywords: Nonlinear model reduction, Balanced realization, variational system, controllability and observability Gramians, extended differential singular values

1. INTRODUCTION

Model reduction is the method of designing a lower-dimensional copy of the original high-dimensional model of the dynamical system which mimics the dominant behaviour of the actual system with considerable accuracy. Balanced truncation is one of the well known and effective *model reduction* techniques which was first proposed in Moore (1981). A thorough exposition of different model reduction techniques for linear systems can be found in Antoulas (2005). Recently an extension of generlized balanced truncation has been put forward in Borja et al. (2022) for continuous-time LTI systems which can provide tighter priori error bounds and brings on the flexibility of preserving structures such as a port-Hamiltonian structure, an electrical network structure, etc. For nonlinear systems balancing has been introduced for the first time in Scherpen (1993). Afterwards there have been several developments in different types of balancing, minimality considerations, association with nonliear Hankel operator in e.g. Verriest and Gray (2000), Fujimoto and Scherpen (2010), Gray and Verriest (2006), Lall et al. (2002). Recently a new approach of nonlinear model reduction has been introduced in Kawano and Scherpen (2017a) which is called *differential balanced truncation*. Differential balancing is basically performing balancing in the contraction framework where a nonlinear system and its variational dynamics have been considered together, which is being called a prolonged system. Generalized differential balancing is computationally tractable way of nonlinear balancing for systems with constant input vector-fields and linear output vectors. Generalized differential balanced truncation has been proposed in Kawano and Scherpen (2015) to find a reduced order model for this kind of nonlinear systems with a prior error bound and stability guarantees. In this work we have extended this idea to extended differential balancing which can provide a less conservative prior error bound in comparison with generalized differential balanced truncation. Apart from that it brings on possibility of preserving certain structures and properties of the nonlinear dynamical system.

On the other hand, structure-preserving model reduction methods for port-Hamiltonian systems have been studied via various approaches, e.g. via balanced truncation Lopezlena et al. (2003),Fujimoto (2008),Kawano and Scherpen (2018), moment matching Polyuga and van der Schaft (2010), Ionescu and Astolfi (2013b), Ionescu and Astolfi (2013a), Kalman decomposition Scherpen and van der Schaft (2008), and Projection-based approach such as Proper Orthogonal Decomposition Chaturantabut et al. (2016). In Lopezlena et al. (2003), balanced truncation is performed via supply and storage functions and the port-Hamiltonian structure is preserved under specific conditions. In Fujimoto (2008), it has been shown that the pH structure is preserved if the Hamiltonian is either identical to so-called weighted controllability function or weighted observability function. In this work, we extend the notion of traditional nonlinear balancing to generalized nonlinear balancing. We define generalized controllability and observability functions which are solutions of two nonlinear partial-differential inequalities and can provide bounds to the traditional controllability and observability functions respectively. Moreover, these generalized functions bring forth the flexibility to propose a balanced truncation approach while preserving the port-Hamiltonian structure in the reduced-order model as well.

2. GENERALIZED DIFFERENTIAL GRAMIANS

Consider the nonlinear system

$$\Sigma : \begin{cases} \dot{x} = f(x) + Bu, \\ y = Cx, \end{cases} \tag{1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$. The function $f : \mathbb{R}^n \to \mathbb{R}^n$ is of class C^1, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{p\times n}$. Let $\phi(t, x_0, u)$ denote the solution $x(t)$ of the system (1) at time $t \in \mathbb{R}_+$ starting from $x_0 \in \mathbb{R}^n$ with input $u(t) \in \mathbb{R}^m$.

Now, we consider the variational system associated with the nonlinear system as

$$d\Sigma : \begin{cases} \delta\dot{x} = \dfrac{\partial f(x)}{\partial x}\delta x + B\delta u, \\ \delta y = C\delta x, \end{cases} \tag{2}$$

where $\delta x(t) \in \mathbb{R}^n$, $\delta u(t) \in \mathbb{R}^m$ and $\delta y(t) \in \mathbb{R}^p$ denote the state, input and output of the variational system respectively. The system (1) together with (2) is called the prolonged system.

The generalized differential controllability and generalized differential observability Gramian are defined as the solutions $P \succ 0$, $Q \succ 0$ of the following differential Lyapunov inequalities

$$\frac{\partial f(x)}{\partial x}P + P\frac{\partial^\top f(x)}{\partial x} + BB^\top \preceq -\epsilon P, \tag{3}$$

$$Q\frac{\partial f(x)}{\partial x} + \frac{\partial^\top f(x)}{\partial x}Q + C^\top C \preceq -\epsilon Q, \tag{4}$$

for all $x \in \mathbb{R}^n$, $\epsilon > 0$. This is slightly different from the original definition of generalized differential Gramians Kawano and Scherpen (2017b),Kawano and Scherpen (2015) where $\epsilon = 0$ is assumed. However, $\epsilon > 0$ guarantees stability properties of the system (1) based on the generalized differential Gramians Kawano (2022). If the solutions exist then the function $\bar{L}_{\mathcal{P}}(\delta x_0) := \frac{1}{2}\delta x_0^\top P^{-1}\delta x_0$ is said to be the generalized differential controllabilty function and the function $\bar{L}_{\mathcal{O}}(\delta x_0) := \frac{1}{2}\delta x_0^\top Q\delta x_0$ is said to be the generalized differential observability function. The solutions of (3) and (4) are not unique and consequently the generalized differential functions are not unique. However, they provide bounds to differential energy functions as follows $\bar{L}_{\mathcal{C}}(\delta x_0) \leq L_{\mathcal{C}}(\delta x_0, x_0)$, $\bar{L}_{\mathcal{O}}(\delta x_0) \geq L_{\mathcal{O}}(\delta x_0, x_0)$, where $x_0 \in \mathbb{R}^n$, $\delta x_0 \in \mathbb{R}^n$, $L_{\mathcal{C}}(\delta x_0, x_0)$ and $L_{\mathcal{O}}(\delta x_0, x_0)$ are the differential controllability function and the differential observability function respectively as defined in Kawano and Scherpen (2015).

3. EXTENDED DIFFERENTIAL GRAMIANS

Extended Gramians have been defined for discrete-time LTI systems in Sandberg (2010) and for continuous-time systems in Scherpen and Fujimoto (2018),Borja et al. (2022). In this section, we use a similar notion to define extended differential Gramians for nonlinear systems. Before proceeding any further, we have the following standing assumption.

Assumption 1. The Jacobian $\partial f(x)/\partial x$ is globally bounded with respect to its argument i.e. $|\frac{\partial f_i}{\partial x_j}| < \infty$, $1 \leq i \leq n$, $1 \leq j \leq n$ for all $x \in \mathbb{R}^n$.

Consider solutions of (3) and (4) are positive definite. Now, we can define $\check{P} := P^{-1}$, where $\check{P}$ is positive definite. For the ease of readability we define

$$\begin{aligned} X_o(x) &:= -Q\frac{\partial f(x)}{\partial x} - \frac{\partial^\top f(x)}{\partial x}Q - C^\top C - \epsilon Q, \\ X_c(x) &:= -\check{P}\frac{\partial f(x)}{\partial x} - \frac{\partial^\top f(x)}{\partial x}\check{P} - \check{P}BB^\top\check{P} - \epsilon\check{P}. \end{aligned} \tag{5}$$

Now, let us consider the following two time-varying LMIs,

$$\begin{bmatrix} X_o(x) & Q - (\alpha I_n + \frac{\partial^\top f}{\partial x})S \\ Q - S^\top(\alpha I_n + \frac{\partial f}{\partial x}) & (S + S^\top) \end{bmatrix} \succeq 0 \tag{6}$$

and

$$\begin{bmatrix} -\check{P}\frac{\partial f}{\partial x} - \frac{\partial^\top f}{\partial x}\check{P} - \epsilon\check{P} & -\check{P} + (\beta I_n + \frac{\partial^\top f}{\partial x})T & -2\check{P}B \\ \check{P} - T^\top(\beta I_n + \frac{\partial f}{\partial x}) & T + T^\top & 2T^\top B \\ -2B^\top\check{P} & 2B^\top T & 4I_m \end{bmatrix} \succeq 0 \tag{7}$$

where $X_o(x)$ is as defined in (5) and $\alpha, \beta > 0$. We then call the solutions (Q, S, α) of (6) and $(\check{P}, T, \beta)$ of (7) as extended differential observability Gramian and inverse of the extended differential controllability Gramian for the variational system (2) respectively.

Theorem 1. Assume $X_o(x) \succ 0$ for all $x \in \mathbb{R}^n$, then the inequality (4) has a solution Q for all $x \in \mathbb{R}^n$ if and only if (6) has a solution (Q, S, α) with $Q \succ 0$, $S = S^\top \succ 0$ and α large enough for all $x \in \mathbb{R}^n$.

Theorem 2. Assume $X_c(x) \succ 0$ for all $x \in \mathbb{R}^n$, then the inequality (3) has a solution P for all $x \in \mathbb{R}^n$ if and only if (7) has a solution $(\check{P}, T, \beta)$ with $\check{P} \succ 0$, $T = T^\top \succ 0$ and β large enough for all $x \in \mathbb{R}^n$.

Now, the nonlinear system (1) is said to be extended differentially balanced if there exists an invertible matrix $W_e \in \mathbb{R}^{n\times n}$ which transforms the system (1) and (2) into a realization where

$$W_e^\top T^{-1} W_e = W_e^{-1} S W_e^{-\top} = \Lambda_{ST}, \tag{8}$$

such that

$$\Lambda_{ST} = \mathsf{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_n\},$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$.

Remark 1. It can be shown in a similar fashion as for linear systems that there always exists an invertible matrix $W_e \in \mathbb{R}^{n\times n}$ which satisfies (8). Moreover, it can be observed that $T^{-1}S = W_e^{-1}\Lambda_{ST}^2 W_e$ which implies that $\sigma_i (i = 1, 2, \cdots, n)$ are the singular values of $T^{-1}S$. We call them the **extended differential singular values** of the prolonged nonlinear system.

4. MODEL REDUCTION AND THE ERROR BOUND

One of the essential features of balanced truncation for linear systems is that it can provide an apriori error bound for the reduction. For linear systems this error bound can be computed through frequency domain analysis of the actual and reduced order models. On the other hand a prior bound could not be provided for balanced truncation of nonlinear systems. However, as generalized differential balancing and extended differential balancing exploit the variational dynamics of the nonlinear system (which is essentially the linearization along the trajectories of the nonlinear system), an apriori error bound can be rendered for the reduced order nonlinear model.

To proceed further let us consider the nonlinear system and the associated variational dynamics in balanced coordinates as

$$\bar{\Sigma} : \begin{cases} \dot{\bar{x}} = \bar{f}(\bar{x}) + \bar{B}u \\ y = \bar{C}\bar{x} \end{cases} \tag{9}$$

$$d\bar{\Sigma} : \begin{cases} \delta\dot{\bar{x}} = \dfrac{\partial \bar{f}(\bar{x})}{\partial \bar{x}}\delta\bar{x} + \bar{B}\delta u \\ \delta y = \bar{C}\delta\bar{x} \end{cases} \tag{10}$$

where,

$$\begin{aligned} &\bar{f}(\bar{x}) := W_e^{-1} f(W_e\bar{x}), \quad \frac{\partial \bar{f}(\bar{x})}{\partial \bar{x}} := W_e^{-1}\frac{\partial f(x)}{\partial x}W_e, \\ &\bar{B} := W_e^{-1}B, \quad \bar{C} := CW_e, \quad x = W_e\bar{x}. \end{aligned} \tag{11}$$

We can split $\bar{x}$ as

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix},$$

where $\bar{x}_1 \in \mathbb{R}^r$ is the part of the state to be kept and $\bar{x}_2 \in \mathbb{R}^{n-r}$, is the part of the state to be truncated while performing the model reduction. Similarly, we can split

$$\bar{f}(\bar{x}) = \begin{bmatrix} \bar{f}_1(\bar{x}_1, \bar{x}_2) \\ \bar{f}_2(\bar{x}_1, \bar{x}_2) \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} \bar{C}_1 & \bar{C}_2 \end{bmatrix} \quad (12)$$

with $\bar{f}_1 : \mathbb{R}^n \to \mathbb{R}^r$, $\bar{f}_2 : \mathbb{R}^n \to \mathbb{R}^{n-r}$, $\bar{B}_1 \in \mathbb{R}^{r\times m}$, $\bar{B}_2 \in \mathbb{R}^{(n-r)\times m}$, $\bar{C}_1 \in \mathbb{R}^{p\times r}$ and $\bar{C}_2 \in \mathbb{R}^{p\times(n-r)}$.

Hence, the truncation of the state component $\bar{x}_2$ leads to the the reduced-order model as follows

$$\hat{\Sigma} : \begin{cases} \dot{\hat{x}} = \hat{f}(\hat{x}) + \hat{B}u \\ \hat{y} = \hat{C}\hat{x}, \end{cases} \quad (13)$$

where,

$$\hat{x} := \bar{x}_1, \quad \hat{f} := \bar{f}_1, \quad \hat{B} := \bar{B}_1, \quad \hat{C} := \bar{C}_1.$$

Assumption 2. The drift vector field of the nonlinear system is an odd function of the state vector, i.e. $f(x) = -f(-x)$.

This assumption is used to provide an apriori error bound for the reduced order model. Though the assumption seems to be quite conservative in nature, odd nonlinear functions for the drift occur in several physical systems e.g. mass-spring-damper systems with nonlinear springs, a nonlinear pendulum, mechanical systems with frictional, backlash nonlinearities, electronic circuits with nonlinear resistors, etc. In addition to that standard static nonlinearities such saturations can be modeled by a hyperbolic tangent function which is odd in nature as well.

Theorem 3. Suppose the system (1) is balanced with the extended differential observability Gramian (Q, S, α) and inverse of extended differential controllability Gramian $(\check{P}, T, \beta)$ as defined in (6) and (7) respectively. If the system is in the balanced coordinates, i.e.

$$S = T^{-1} = \Lambda_{ST} = \mathsf{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_n\},$$

and the drift vector field is odd in nature, then the output of the actual model and output of reduced order model satisfy

$$||y - \hat{y}||_2 \leq 2 \sum_{j=r+1}^{n} \sigma_j ||u||_2$$

where $\sigma_r \gg \sigma_{r+1}$ and $\alpha = \beta$.

5. GENERALIZED ENERGY FUNCTIONS FOR NONLINEAR SYSTEMS

In this section we introduce the generalized observability and generalized controllability functions. These generalized energy functions provide bounds to the original energy functions as is evident from the following proposition.

Theorem 4. If $\tilde{L}_o(x) > 0$ is a smooth solution of

$$\frac{\partial \tilde{L}_o(x)}{\partial x}(x) f(x) + \frac{1}{2} h^\top(x) h(x) \leq 0, \quad \tilde{L}_o(0) = 0, \quad (14)$$

and if the system is locally stable in a neighbourhood of the origin, then $\tilde{L}_o(x_0)$ is called the generalized observability function which satisfies

$$L_o(x_0) \leq \tilde{L}_o(x_0). \quad (15)$$

Moreover, if $\tilde{L}_c(x) > 0$ is a smooth solution of

$$\frac{\partial \tilde{L}_c(x)}{\partial x} f(x) + \frac{1}{2} \frac{\partial \tilde{L}_c(x)}{\partial x} g(x) g^\top(x) \frac{\partial^\top \tilde{L}_c(x)}{\partial x} \leq 0, \tilde{L}_c(0) = 0, \quad (16)$$

and if there exists an anti-stabilizing input $u(x)$ such that $x(-\infty) = 0$ and $x(0) = x_0$, then then $\tilde{L}_c(x_0)$ is called the generalized controllability function which satisfies

$$L_c(x_0) \geq \tilde{L}_c(x_0). \quad (17)$$

6. GENERALIZED BALANCING FOR NONLINEAR PORT-HAMILTONIAN SYSTEMS

Consider an input-state-output nonlinear port-Hamiltonian system

$$\Sigma_{PH} : \begin{cases} \dot{x} = (J(x) - R(x)) \dfrac{\partial \mathcal{H}(x)}{\partial x} + g(x)u, \\ y = g^\top(x) \dfrac{\partial \mathcal{H}(x)}{\partial x}, \end{cases} \quad (18)$$

where $t \in \mathbb{R}$, the state $x(t) \in \mathbb{R}^n$, the input $u(t) \in \mathbb{R}^m$, and the output $y(t) \in \mathbb{R}^m$. $J(x) = -J^\top(x)$ and $R(x) = R^\top(x) \succeq 0$ represent the interconnection and dissipation matrix respectively. $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}$, $\mathcal{H}(x) \succ 0$ is the Hamiltonian function which represents the total energy of the system. To simplify the notation, we define $F(x) := (J(x) - R(x))$. Without loss of generality let us consider that $\frac{\partial \mathcal{H}}{\partial x}(0) = 0$ and $\frac{\partial^2 \mathcal{H}}{\partial x^2}(0) \succ 0$. We also consider that the system is asymptotically stable on a neighbourhood of the origin.

Assumption 3. For a nonlinear port-Hamiltonian system (18), assume the following holds

- 0 is an asymptotically stable equilibrium of $F(x)\frac{\partial \mathcal{H}(x)}{\partial x}$ on some neighbourhood W of 0.
- The linearized system at the origin is asymptotically stable.
- (14) and (16) have smooth solutions on W.
- $\frac{\partial^2 \tilde{L}_c}{\partial x^2}(0) \succ 0$, $\frac{\partial^2 \tilde{L}_o}{\partial x^2}(0) \succ 0$ and $\frac{\partial^2 \mathcal{H}}{\partial x^2}(0) \succ 0$.

Assumption 4. There exists $\tilde{L}_c(x)$, $\tilde{L}_o(x) > 0$ such that the eigenvalues of $\frac{\partial^2 \tilde{L}_c}{\partial x^2}(0)^{-1} \frac{\partial^2 \tilde{L}_o}{\partial x^2}(0)$ as well as the eigenvalues of $\frac{\partial^2 \mathcal{H}}{\partial x^2}(0)^{-1} \frac{\partial^2 \tilde{L}_c}{\partial x^2}(0)^{-1} \frac{\partial^2 \tilde{L}_o}{\partial x^2}(0)$ are distinct.

Theorem 5. Suppose that Assumption 3 and 4 hold. Then there exists a transformation $\bar{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$ on a neighbourhood U of the origin such that $x = \bar{\Phi}(\bar{z}), \bar{\Phi}(0) = 0$ which converts the system into a new realization where the following holds

$$\begin{aligned} \tilde{L}_c(\bar{\Phi}(\bar{z})) &= \frac{1}{2} \sum_{i=1}^{n} \frac{\bar{z}_i^2}{\bar{\sigma}_i(\bar{z}_i)}, \\ \tilde{L}_o(\bar{\Phi}(\bar{z})) &= \frac{1}{2} \sum_{i=1}^{n} \bar{z}_i^2 \bar{\sigma}_i(\bar{z}_i), \\ \mathcal{H}(\bar{\Phi}(\bar{z})) &= \frac{1}{2} \sum_{i=1}^{n} z_i^2 \frac{\bar{\eta}_i(\bar{z}_i)}{\bar{\sigma}_i(\bar{z}_i)}, \end{aligned} \quad (19)$$

where $\bar{\sigma}_1(\bar{z}_1) \geq \bar{\sigma}_2(\bar{z}_2) \geq \cdots \geq \bar{\sigma}_n(\bar{z}_n)$ and $\bar{\eta}_1(\bar{z}_1) \geq \bar{\eta}_2(\bar{z}_2) \geq \cdots \geq \bar{\eta}_n(\bar{z}_n)$ are smooth functions.

Now, we can split $\bar{z}$ into two parts as $\bar{z} = [\bar{z}_r^\top, \bar{z}_t^\top]^\top$, where $\bar{z}_r = [\bar{z}_1, \bar{z}_2, \cdots, \bar{z}_k]^\top \in \mathbb{R}^n$ and $\bar{z}_t = [\bar{z}_{k+1}, \bar{z}_{k+2}, \cdots, \bar{z}_n]^\top \in \mathbb{R}^{n-k}$. Similarly, we can split

$$J_{\bar{z}} = \begin{bmatrix} J_{\bar{z},rr}(\bar{z}_r,\bar{z}_t) & J_{\bar{z},rt}(\bar{z}_r,\bar{z}_t) \\ -J_{\bar{z},rt}^{\top}(\bar{z}_r,\bar{z}_t) & J_{\bar{z},tt}(\bar{z}_r,\bar{z}_t) \end{bmatrix},$$
$$R_{\bar{z}} = \begin{bmatrix} R_{\bar{z},rr}(\bar{z}_r,\bar{z}_t) & R_{\bar{z},rt}(\bar{z}_r,\bar{z}_t) \\ R_{\bar{z},rt}(\bar{z}_r,\bar{z}_t) & R_{\bar{z},tt}(\bar{z}_r,\bar{z}_t) \end{bmatrix},$$
$$g_{\bar{z}} = \begin{bmatrix} g_{\bar{z},r}(\bar{z}_r,\bar{z}_t) \\ g_{\bar{z},t}(\bar{z}_r,\bar{z}_t) \end{bmatrix},$$

where $J_{\bar{z},rr}(\bar{z}_r,\bar{z}_t)$ and $J_{\bar{z},tt}(\bar{z}_r,\bar{z}_t)$ are skew-symmetric, $R_{\bar{z},rr}(\bar{z}_r,\bar{z}_t)$ and $R_{\bar{z},tt}(\bar{z}_r,\bar{z}_t)$ are symmetric and positive semidefinite. In addition to that we can split the Hamiltonian of the system in two parts as follows

$$\mathcal{H}_{\bar{z}}(\bar{z}) = \mathcal{H}_{\bar{z}}(\bar{z}_r,0) + \mathcal{H}_{\bar{z}}(0,\bar{z}_t). \tag{20}$$

Theorem 6. Consider a continuous-time nonlinear input-state-output port-Hamiltonian system Σ_{PH} as in (18). Suppose that Assumption 3 and Assumption 4 are satisfied and we obtain a balanced realization of the system as in (19). Then a reduced-order model of (18) can be represented as follows

$$\begin{aligned} \dot{\bar{z}}_r &= (J_{\bar{z},rr}(\bar{z}_r,0) - R_{\bar{z},rr}(\bar{z}_r,0))\frac{\partial \mathcal{H}_{\bar{z}}(\bar{z}_r,0)}{\partial \bar{z}_r} + g_{\bar{z},r}(\bar{z}_r,0)u, \\ y_r &= g_{\bar{z},r}^{\top}(\bar{z}_r,0)\frac{\partial \mathcal{H}_{\bar{z}}(\bar{z}_r,0)}{\partial \bar{z}_r}, \end{aligned} \tag{21}$$

which is also a port-Hamiltonian system with the Hamiltonian $\mathcal{H}_{\bar{z}}(\bar{z}_r,0)$ representing the total energy of the reduced model. Moreover, $\tilde{L}_o(\bar{\Phi}(\bar{z}_r,0))$ and $\tilde{L}_c(\bar{\Phi}(\bar{z}_r,0))$ satisfy (14) and (16) respectively for the reduced order system.

We will include the corresponding proofs, further details and the application of the theoretical results via an illustrative example in the presentation.

ACKNOWLEDGEMENT

This publication is part of the project Digital Twin P18-03 project 1 of the research program Perspectief which is (partly) financed by the Dutch Research Council (NWO).

REFERENCES

Antoulas, A. (2005). *Approxiation of Large-scale Dynamical Systems*. Society for Industrial and Applied Mathematics.

Borja, P., Scherpen, J.M.A., and Fujimoto, K. (2022). Extended balancing of continuous-time LTI systems : a structure-preserving approach. *IEEE Transactions on Automatic Control(early access)*. doi: https://doi.org/10.1109/TAC.2021.3138645.

Chaturantabut, S., Beattie, C., and Gugercin, S. (2016). Structure-preserving model reduction for nonlinear port-Hamiltonian systems. *SIAM journal on Scientific Computing*, 38(5), B837–B865.

Fujimoto, K. and Scherpen, J.M.A. (2010). Balanced realization and model reduction for nonlinear systems based on singular value analysis. *SIAM journal on Control and Optimization*, 48(7), 4591–4623.

Fujimoto, K. (2008). Balanced realization and model order reduction for port-Hamiltonian systems. *Journal of System Design and Dynamics*, 2(3), 694–702.

Gray, W.S. and Verriest, E.I. (2006). Balancing near stable invariant manifolds. *Automatica*, 42, 654–659.

Ionescu, T.C. and Astolfi, A. (2013a). Families of moment matching based, structure preserving approximations for linear port-Hamiltonian systems. *Automatica*, 49(8), 2424–2434.

Ionescu, T.C. and Astolfi, A. (2013b). Moment matching for nonlinear port-Hamiltonian and gradient systems. In *Proceedings of 9th IFAC Symposium on Nonlinear Control Systems*, 395–399.

Kawano, Y. (2022). Controller reduction for nonlinear systems by generalized differential balancing. *IEEE Transactions on Automatic Control(early access), doi: 10.1109/TAC.2021.3124980*, 67(11). doi: 10.1109/TAC.2021.3124980.

Kawano, Y. and Scherpen, J.M.A. (2015). Model reduction by generalized differential balancing. In M.K. Camlibel, A.A. Julius, R. Pasumarthy, and M.A. Scherpen Jacquelien (eds.), *Mathematical Control theory I*, chapter 19, 349–362. Springer-verlag.

Kawano, Y. and Scherpen, J.M.A. (2017a). Model reduction by differential balancing based on nonlinear hankel operators. *IEEE Transactions on Automatic Control*, 62(7), 3293–3308.

Kawano, Y. and Scherpen, J.M.A. (2017b). Model reduction by differential balancing based on nonlinear Hankel operators. *IEEE Transactions on Automatic Control*, 62(7), 3293–3308.

Kawano, Y. and Scherpen, J.M.A. (2018). Structure preserving truncation of nonlinear port-Hamiltonian systems. *IEEE Transactions on Automatic Control*, 63(12), 4286–4293.

Lall, S., Marsden, J.E., and Glavaski, S. (2002). A subspace approach to balanced truncation for model reduction of nonlinear control systems. *International Journal of Robust and Nonlinear Control*, 12, 519–535.

Lopezlena, R., Scherpen, J.M.A., and Fujimoto, K. (2003). Energy-storage balanced reduction of port-Hamiltonian systems. In *Proceedings of 2nd IFAC workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control*, volume 36, 69–74.

Moore, B. (1981). Principal compoent analysis in linear systems : Controllability, observability and model reduction. *IEEE Transactions on Automatic Control*, 26(1), 17–32.

Polyuga, R.V. and van der Schaft, A.J. (2010). Structure-preserving model reduction of port-hamiltonian systems by moment matching at infinity. *Automatica*, 46(4), 665–672.

Sandberg, H. (2010). An extension to balanced truncation with application to structured model reduction. *IEEE Transactions on Automatic Control*, 55(4), 1038–1043.

Scherpen, J.M.A. (1993). Balancing for nonlinear systems. *Systems and Control Letters*, 21(2), 143–153.

Scherpen, J.M.A. and Fujimoto, K. (2018). Extended balanced truncation for continuous time LTI systems. In *the Proceedings of European Control Conference (ECC)*, 2611–2615.

Scherpen, J.M.A. and van der Schaft, A.J. (2008). A structure-preserving minimal representation of a nonlinear port-Hamiltonian system. In *Proceedings of 47th IEEE Conference on Decision and Control*, 4885–4890.

Verriest, E.I. and Gray, W.S. (2000). Flow balancing nonlinear systems. In *14th International Symposium on Mathematical Theory of Networks and Systems*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Semialgebraic Convex Bodies

Chiara Meroni *

* *Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany (e-mail: chiara.meroni@mis.mpg.de)*

Abstract
Convex Algebraic Geometry lives at the intersection of Convex Geometry, Optimization, Algebraic Geometry and Real Algebra. Classically, convex geometry has been studied from an analytical point of view. Here, we approach it using tools from real and complex algebraic geometry, with a focus on semialgebraic convex bodies, beyond polytopes.

Keywords: Convex Algebraic Geometry, Algebraic boundary, Convex bodies, Zonoids, Semialgebraic geometry.

MSC classes: 52A99, 14P10, 52A21.

What is convex algebraic geometry? By this name we refer to the study of convex geometry from the point of view of algebraic geometry and real algebra. This approach has its origin in the theory of polytopes, connected to linear algebra and combinatorics. From there, it is natural to go beyond linear algebra and enter the world of *nonlinear algebra* (Michałek and Sturmfels (2021)).

In this setting we study the family of semialgebraic convex bodies. They are convex, compact, non-empty, semialgebraic (see Bochnak et al. (2013)) subsets of some Euclidean space. These objects were first studied in the context of semidefinite and polynomial optimization; in particular, we refer to Lasserre (2009) and Blekherman et al. (2013). Convex geometry is classically interested in analytical and functional aspects of convex bodies, and these behave well with respect to semialgebraicity. Indeed, given a convex body $K \subset \mathbb{R}^n$, the following are equivalent:

- K is semialgebraic;
- the support function of K is a semialgebraic function;
- the radial function of K is a semialgebraic funciton;
- the dual/polar body of K is semialgebraic.

Moving towards algebraic geometry and convex geometry, the object that better encodes this interaction is the *algebraic boundary*. Let $K \subset \mathbb{R}^n$ be a convex body. Its *algebraic boundary* is the smallest variety that contains the topological boundary of K. In other words, it is the closure of the topological boundary with respect to the Zariski topology. In this way, we associate a variety to a convex body. We can study such a variety using tools from algebraic geometry in order to get information about K. For instance, the algebraic boundary detects the semialgebraicity of a convex body: K is semialgebraic if and only if its algebraic boundary is an algebraic hypersurface. Polytopes are an example of semialgebraic convex bodies, and their algebraic boundary is the hyperplane arrangement defined by the facets. One hopes to extend notions and techniques from the theory of polytopes to semialgebraic convex bodies. Some examples are Plaumann et al. (2021), where the authors develop a broad definition of an f-vector, and Saunderson and Chandrasekaran (2020), that discusses a generalization of the neighborliness for non-polyhedral convex cones.

One can generalize polytopes in many ways, in order to obtain classes of semialgebraic convex bodies. The family of *spectrahedra* is one option (Ramana and Goldman (1995)). They arise as the intersection of the cone of positive semidefinite matrices with a linear subspace. Spectrahedra are relevant in optimization because they are the feasible regions of semidefinite programming. Their study is intimately related to the study of matrices and determinantal varieties. Another direction is that of the *convex hull of a variety* (Ranestad and Sturmfels (2011, 2012)). Understanding the boundary of a convex hull is a difficult task in general. However, algebraic geometry gives the answer in the case of convex hulls of varieties, as stated in (Ranestad and Sturmfels, 2011, Theorem 1.1). Such a formula describes the components of the algebraic boundary of K, where K is the convex hull of a smooth compact real algebraic variety in $\mathbb{R}^n$.

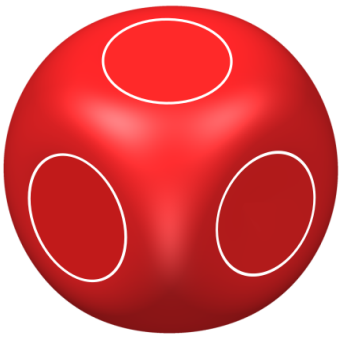


Figure 1. The Minkowski sum of three 2-dimensional discs.

Not all convex bodies are semialgebraic. For instance, *zonoids* have a non-empty intersection with the set of semialgebraic convex bodies, but are not contained in it. Hence, the immediate question is: which zonoids are semialgebraic? This lies in the context of the Zonoid Problem

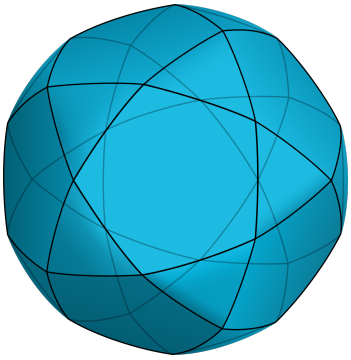


Figure 2. The intersection body of the icosahedron.

(Blaschke (1923); Bolker (1971)). This problem is very hard to tackle (see Weil (1977, 1982)): restricting to the subclass of semialgebraic convex bodies would potentially make it easier. In Gesmundo and Meroni (2022), we investigate a class of semialgebraic zonoids called *discotopes*. They are Minkowski sums of finitely many discs, see Figure 1. We study their algebraic boundary, in order to be able to characterize them. The beauty and the strength of this problem is that it can be approached from many points of view: algebraic geometry, as in our work, measure theory, random geometry.

Many areas of convex geometry investigate objects that are defined starting from a convex body. A goal is to understand how semialgebraic geometry behaves with respect to these constructions.

The notion of *fiber polytope* was introduced by Billera and Sturmfels (1992). It encodes a number of combinatorial features about the polytopes involved. Notably, the same definition works for more general convex bodies. In Mathis and Meroni (2021), we focused on this generalization in the following setting. Let K be a convex body in $\mathbb{R}^{n+m}$ and $\pi : \mathbb{R}^{n+m} \to V$ be the orthogonal projection onto a subspace $V \subset \mathbb{R}^{n+m}$ of dimension n. The fiber body of K with respect to π is the *average* of the fibers of K under this projection:

$$\Sigma_\pi K = \int_{\pi(K)} \left(K \cap \pi^{-1}(x)\right) dx \subset \mathbb{R}^m,$$

where this is a Minkowski integral. Among other properties, we prove that the fiber body of a zonoid, with respect to any projection, is again a zonoid. On the other hand, this construction does not behave well with semialgebraicity: we provide an example of a semialgebraic convex body, the dice in Figure 1, having a non-semialgebraic fiber body.

Another interesting construction that plays a central role in geometric tomography (see Gardner (2006)) is that of the *intersection body*. Let $K \subset \mathbb{R}^n$ be a full dimensional convex body. Its intersection body is the set $IK = \{x \in \mathbb{R}^n \,|\, \rho_K(x) \geq 1\}$, for the radial function

$$\rho_K(x) = \frac{1}{\|x\|} \operatorname{vol}\left(K \cap x^\perp\right),$$

where $x^\perp$ is the hyperplane orthogonal to x. In Berlow et al. (2022) we examined the case when K is a polytope; Figure 2 shows the intersection body of a icosahedron. Our main contribution states that the intersection body of a polytope is a semialgebraic set. It can be associated to a zonotope whose face structure reflects that of the intersection body. We provide an algorithm for computing intersection bodies of polytopes. Its implementation is available at `https://mathrepo.mis.mpg.de/intersection-bodies`, together with interactive three-dimensional models that highlight interesting features.

REFERENCES

Berlow, K., Brandenburg, M.C., Meroni, C., and Shankar, I. (2022). Intersection bodies of polytopes. *Beiträge zur Algebra und Geometrie*.

Billera, L.J. and Sturmfels, B. (1992). Fiber polytopes. *Annals of Mathematics*, 135(3), 527–549.

Blaschke, W. (1923). *Vorlesungen über Differentialgeometrie. II.* Springer-Verlag, Berlin.

Blekherman, G., Parrilo, P.A., and Thomas, R.R. (eds.) (2013). *Semidefinite Optimization and Convex Algebraic Geometry*, volume 13 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM); Mathematical Optimization Society, Philadelphia, PA.

Bochnak, J., Coste, M., and Roy, M.F. (2013). *Real algebraic geometry*, volume 36. Springer Science & Business Media.

Bolker, E.D. (1971). The zonoid problem. *Amer. Math. Monthly*, 78(5), 529–531.

Gardner, R.J. (2006). *Geometric Tomography*, volume 58 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, New York, second edition.

Gesmundo, F. and Meroni, C. (2022). The geometry of discotopes. *Le Matematiche*, 77, 143–171.

Lasserre, J.B. (2009). *Moments, positive polynomials and their applications*, volume 1 of *Series on Optimization and Its Applications*. World Scientific.

Mathis, L. and Meroni, C. (2021). Fiber Convex Bodies. *arXiv:2105.12406*.

Michałek, M. and Sturmfels, B. (2021). *Invitation to nonlinear algebra*, volume 211. American Mathematical Soc.

Plaumann, D., Sinn, R., and Wesner, J.L. (2021). Families of faces and the normal cycle of a convex semi-algebraic set. *arXiv:2104.13306*.

Ramana, M. and Goldman, A.J. (1995). Some geometric results in semidefinite programming. *Journal of Global Optimization*, 7(1), 33–50.

Ranestad, K. and Sturmfels, B. (2011). The convex hull of a variety. In P. Brändén, M. Passare, and M. Putinar (eds.), *Notions of Positivity and the Geometry of Polynomials*, 331–344. Springer Verlag, Basel.

Ranestad, K. and Sturmfels, B. (2012). On the convex hull of a space curve. *Advances in Geometry*, 12(1), 157–178.

Saunderson, J. and Chandrasekaran, V. (2020). Terracini convexity. *arXiv:2010.00805*.

Weil, W. (1977). Blaschkes Problem der lokalen Charakterisierung von Zonoiden. *Arch. Math.*, 29(1), 655–659.

Weil, W. (1982). Zonoide und verwandte Klassen konvexer Körper. *Monatsh. Math.*, 94(1), 73–84.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A turnpike result for optimal boundary control of gas pipeline flow

Martin Gugat * **Michael Herty** **

* *Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department of Data Science, Lehrstuhl für Dynamics, Control and Numerics (Alexander von Humboldt-Professur), Cauerstr. 11, 91058 Erlangen, Germany (e-mail: martin.gugat@fau.de).*
** *Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany (e-mail: herty@igpm.rwth-aachen.de)*

Abstract: The operation of gas pipeline networks leads to problems of optimal boundary control for systems governed by the isothermal Euler equations. In this contribution we consider a problem of optimal Dirichlet control with an objective function of integral type that is given as the sum of a tracking term for a desired stationary state and the corresponding control cost. We study the well-posedness and exact controllability properties of the system. We study regular solutions that generate a field of non-intersecting characteristic curves without rarefaction fan that transport the information about the state. We also consider the exact controllability properties of the system. We define a problem of optimal boundary control and show the existence of a solution. We present an integral turnpike result and a result about the turnpike phenomenon with interior decay for such an optimal control problem.

Keywords: gas pipeline flow, boundary control, optimal control, hyperbolic partial differential equation, turnpike phenomenon, isothermal Euler equations

1. INTRODUCTION

We consider an optimal control problem for the operation of gas pipelines. As a model for gas pipeline flow we consider the isothermal Euler equations (see e.g. Banda et al. (2006), Gugat and Herty (2022))

$$\begin{cases} \rho_t + q_x = 0, \\ q_t + \left(p + \dfrac{q^2}{\rho}\right)_x = -\dfrac{1}{2}\theta \dfrac{q\,|q|}{\rho}. \end{cases} \tag{1}$$

Here ρ denotes the gas density, p the pressure q the mass flow rate and θ is a friction parameter. At the end $x = 0$, the desired pressure $p_d > 0$ is prescribed, $p(t,\, 0) = p_d$. At the other end $x = L$ of the pipe, the flow rate is controlled, $q(t,\, L) = u(t)$. In the optimal control problem, state and control constraints in C^1 enforce the regularity of the generated states. In Gugat and Sokolowski (2022), a similar optimal control problem for gas networks is considered and the existence of an optimal control is shown. Equation (1) is a model for the flow through a horizontal pipeline. For a sloped pipeline, an additional source term that is proportional to the sine of the angle of inclination φ appears in the second equation on the right-hand side, namely $-g\,\rho\,\sin(\varphi)$, where g is the gravitational constant.

⋆ This work was funded by the DFG, TRR 154, *Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks*, projects C03 and C05, Projektnummer 239904186

2. THE SYSTEM

Let a space interval $[0,\, L]$ be given; here $L > 0$ corresponds to the length of the pipe. Let a classical subsonic steady state $p_r(x)$, $q_r(x)$ ($x \in [0,\, L]$) with the constant control u_r and $p_r(0) = p_d$ be given. Assume that for all $x \in [0, L]$ we have $p_r(x) > 0$ and $(p_r,\, q_r) \in (H^2(0,\, L))^2$. For the case of ideal gas where $p = a^2\rho$ with the sound speed $a > 0$, the initial boundary value problem for our system is

$$(\mathbf{S}) \begin{cases} \rho(0,x) = \rho_0(x),\ q(0,x) = q_0(x),\ x \in (0,L), \\ p(t,\,0) = p_r(0),\ \ q(t,\,L) = u(t),\ t \in (0,T), \\ \begin{pmatrix} \rho \\ q \end{pmatrix}_t + \begin{pmatrix} 0 & 1 \\ a^2 - \dfrac{q^2}{\rho^2} & 2\,\dfrac{q}{\rho} \end{pmatrix} \begin{pmatrix} \rho \\ q \end{pmatrix}_x = \begin{pmatrix} 0 \\ -\dfrac{1}{2}\theta \dfrac{q\,|q|}{\rho} \end{pmatrix}. \end{cases}$$

Here $(\rho_0,\, q_0)$ denotes a given initial state.

The theory of semi-global classical solutions (see Li (2010)) asserts that for any given time horizon $T_0 > 0$ there exists a number $\varepsilon(T_0) > 0$ such that for all initial states $(\rho_0,\, q_0)$ that satisfy

$$\max\{\|\rho_0 - \rho_r\|_{C^1([0,\,L])},\, \|q_0 - q_r\|_{C^1([0,\,L])}\} \le \varepsilon(T_0) \tag{2}$$

and are C^1-compatible with p_d as a boundary value at $x = 0$ and all controls with

$$\|u - u_r\|_{C^1([0,\,T_0])} \le \varepsilon(T_0) \tag{3}$$

that are C^1-compatible with the initial state there exists a classical solution of (**S**) on $[0,\, T_0]$ that satisfies the *a priori* estimate

$$\max_{t\in[0,T_0]} \|(p(t,\cdot)-p_r, q(t,\cdot)-q_r)\|_{(C^1([0,L]))^2}$$
$$\leq C_1(T_0)\, \max\{\|\delta\rho\|_{C^1([0,\,L])},\, \|\delta q\|_{C^1([0,\,L])},\, \|\delta u\|_{C^1([0,\,T_0])}\}$$

where we use the notation $\delta\rho = \rho_0 - \rho_r$, $\delta q = q_0 - q_r$, $\delta u = u - u_r$.

Remark 1. Note that a similar result holds with C^1 replaced by H^2, see Bastin and Coron (2016).

To guarantee that a regular solution exists, in our optimal control problem we prescribe the control constraint

$$\|u - u_r\|_{C^1([0,\,T])} \leq \varepsilon_0 \tag{4}$$

with $\varepsilon_0 > 0$ and the C^1-compatibility conditions with the initial state. Moreover, for $T \geq T_0$ we impose the state constraint

$$\max_{t\in[0,T]} \|(p(t,\cdot)-p_r, q(t,\cdot)-q_r)\|_{(C^1([0,L]))^2} \leq C_0\, \varepsilon_0 \tag{5}$$

with $C_0 > 0$.

The constraints (4) and (5) allow to make the time horizon T arbitrarily large without losing the regularity of the solutions. This is possible since for given C^1-initial data (for example at the time T_0) that satisfies $\|(p(T_0,\cdot)-p_r, q(T_0,\cdot)-q_r)\|_{(C^1([0,L]))^2} \leq C_0\, \varepsilon_0$ a classical solution of (**S**) exists for a certain (possibly short) time interval starting at T_0, where a certain minimal length is guaranteed a priori as a function of the C^1-norm of the initial data and the boundary data.

Note that due to the definition of the C^1-norm, the constraints (4), (5) can be written as pointwise constraints for all $t \in [0,\, T]$, which is a form that is well-suited for turnpike analysis. In fact, (4), (5) are equivalent to

$$\begin{cases} |u(t) - u_r| \leq \varepsilon_0,\, t \in [0,\, T], \\ |u_t(t)| \leq \varepsilon_0,\, t \in [0,\, T], \\ |p(t,x) - p_r(x)| \leq C_0\, \varepsilon_0,\, t \in [0,\, T],\, x \in [0, L], \\ |p_x(t,x) - p_r'(x)| \leq C_0\, \varepsilon_0,\, t \in [0,\, T],\, x \in [0, L], \\ |q(t,x) - q_r(x)| \leq C_0\, \varepsilon_0,\, t \in [0,\, T],\, x \in [0, L], \\ |q_x(t,x) - q_r'(x)| \leq C_0\, \varepsilon_0,\, t \in [0,\, T],\, x \in [0, L]. \end{cases}$$

2.1 Exact controllability

System (**S**) is exactly controllable in a sufficiently large time interval:

Theorem 2. Assume that T_0 is sufficiently large and that

$$\max\{\|q_0 - q_r\|_{H_2([0,\,L])},\, \|\rho_0 - \rho_r\|_{H^2([0,\,L])}\} \leq \varepsilon_1 \tag{6}$$

with ε_1 chosen sufficiently small.

Then there exists a control $\hat{u} \in H^2(0, T_0)$ that generates a solution of (**S**) that satisfies the end condition $q(T_0,\, x) = q_r(x)$, $\rho(T_0,\, x) = \rho_r(x)$, $x \in (0, L)$. Moreover, there exists a constant $\hat{C} > 0$ such that $\|\hat{u} - u_r\|_{H^2(0,\,T_0)}$

$$\leq \hat{C} \max\{\|\rho_0 - \rho_r\|_{H^2(0,\,L)}, \|q_0 - q_r\|_{H^2(0,\,L)}\} \tag{7}$$

and

$$\max_{t\in[0,T_0]} \|(p(t,\cdot)-p_r, q(t,\cdot)-q_r)\|_{(H^2([0,L]))^2}$$
$$\leq \tilde{C} \max\{\|\rho_0 - \rho_r\|_{H^2(0,\,L)}, \|q_0 - q_r\|_{H^2(0,\,L)}\}.$$

Proof. The initial data $(q_0,\, p_0)$ and the terminal data $(q_r,\, p_r)$ determine the state on two domains bounded by characteristic curves from different families. Together with boundary data for $x = L$ on $[0,\, T_0]$, these data determine the boundary of a set that consists of a characteristic that starts a $(t, x) = (0, 0)$ until it reaches $x = L$, a piece of $[0,\, T_0] \times \{L\}$ and a characteristic that starts at $x = L$ for some $t < T$ and leads to $(t, x) = (T_0, 0)$. The corresponding boundary data can be obtained by a sufficiently smooth interpolation that does not increase the H^2-norm on the piece of $[0,\, T_0] \times \{L\}$. With these boundary data, the system state is completely determined on $[0,\, T_0] \times [0, L]$. The boundary trace at $x = 0$ yields the desired control. The corresponding a priori inequalities for the system where the stationary state is subtracted yield the inequality (7). The last inequality in the theorem also follows with the H^2 a priori estimate for the state.

Remark 3. In Gugat et al. (2017), a strict H^2-Lyapunov function is introduced for Neumann boundary feedback stabilization of the isothermal Euler equations. In Gugat et al. (2022), the constrained exact boundary controllability of a semilinear model for pipeline gas flow is studied with continuous states.

3. THE OPTIMAL CONTROL PROBLEM

The optimal control problem $\mathbf{P_{dyn}}(T)$ is to find a control function u that is C^1-compatible with $(q_0,\, \rho_0)$ such that the objective function

$$J_T(u) = \|u - u_r\|^2_{H^2(0,\,T)} \tag{8}$$
$$+\|p - p_r\|^2_{L^2([0,T]\times[0,L])} + \|q - q_r\|^2_{L^2([0,T]\times[0,L])}$$

is minimized and (4) and (5) hold, where $(p,\, q)$ denotes the solution of (**S**).

Note that this choice of the objective function mitigates pressure fluctuations and hence is useful to reduce hydrogen embrittlement of steel pipelines during transients. For a description of this effect see Hafsi et al. (2018).

A numerical method that is based upon a DAE approach for optimal control problems in the operation of gas netwrorks including storage reservoirs is presented in Hari et al. (2022).

3.1 Existence of solutions

In Theorem 4 we state that for all $T \geq T_0$ (with T_0 as in Theorem 2), the optimal control problem $\mathbf{P_{dyn}}(T)$ has a solution for suitably chosen problem parameters.

Theorem 4. Assume that $\varepsilon_0 > 0$ is chosen sufficiently small and $C_0 > 0$ is chosen sufficiently large, $(\rho_0,\, q_0) \in (H_2([0,\, L]))^2$ are C^1-compatible with p_d as a boundary value at $x = 0$ and

$$\max\{\|q_0 - q_r\|_{H_2([0,\,L])},\, \|\rho_0 - \rho_r\|_{H^2([0,\,L])}\} \leq \varepsilon_2 \tag{9}$$

with ε_2 chosen sufficiently small. Let $T \geq T_0$ be given.

Then a solution of the dynamic optimal control problem $\mathbf{P_{dyn}}(T)$ does exist.

Proof. Since (9) holds and $T \geq T_0$, Theorem 2 implies that there exists an exact control $u_{exact} \in H^2(0,\, T)$ that steers the system state from the initial state $(p_0,\, q_0)$ in finite time T_0 to $(p_r,\, q_r)$. After this finite time, we extend the control with the constant control value u_r. For u_{exact}, (4) is implied by (7) if ε_2 is chosen sufficiently small. By choosing ε_2 chosen sufficiently small, the a priori inequality implies (5). In this way we obtain a feasible control u_{exact}. Hence the set of admissible controls is non-empty.

We consider a minimizing sequence of feasible controls. Due to the H^2-term in the objective function J_T defined in (8), this sequence is bounded with respect to the H^2-norm hence it contains a subsequence that converges strongly in $C^1([0,\, T])$ to a limit point u^*. Due to the theory of semi-global solutions, the strong convergence implies that also the corresponding subsequence of generated states given by the classical solutions of (**S**) converges strongly to the solution that is generated by the limit point u^*. Moreover, (4) and (5) hold for this limit. Hence u^* is feasible and due to the sequential weak lower semicontinuity of J_T has a minimal value of the objective function. Thus u^* solves $\mathbf{P_{dyn}}(T)$.

3.2 An integral turnpike result

In this section we present our first turnpike result, which is an integral turnpike result. It states that for the optimal control and the generated state the integral in the objective function remains uniformly bounded with respect to T for arbitrarily large T. For a survey on the turnpike property, see the monograph Zaslavski (2019). The exponential turnpike property for optimal control problems in Hilbert spaces is studied in Trélat et al. (2018), see also Grüne et al. (2020), Faulwasser and Kellett (2021) and the references therein.

Theorem 5. Assume that the assumptions of Theorem 4 hold.

Then the solutions of the dynamic optimal control problem $\mathbf{P_{dyn}}(T)$ satisfy a turnpike inequality in the sense that there exists a constant $C_{int} > 0$ such that for all $T > 0$ we have

$$\|u - u_r\|^2_{H^2(0,\,T)} \tag{10}$$
$$+\|p - p_r\|^2_{L^2([0,T]\times[0,L])} + \|q - q_r\|^2_{L^2([0,T]\times[0,L])} \leq C_{int}.$$

Proof. The feasible control $u_{exact} \in H^2(0,\, T)$ from the proof of Theorem 4 is feasible for all $T \geq T_0$ and $J_T(u_{exact})$ is independent of T. Hence with the definition $C_{int} = J_{T_0}(u_{exact}) = J_T(u_{exact})$, the assertion follows.

3.3 The turnpike property with interior decay

In this section we show that the optimal state and the optimal control satisfy a turnpike property with interior decay as it is introduced in Gugat (2021),

Assume that the assumptions of Theorem 4 hold and that $T \geq 2\, T_0$.

Due to (10), we have

$$\|u - u_r\|^2_{H^2(0,\,T)} \leq C_{int}. \tag{11}$$

Moreover, due to the state constraints the state is a classical solution of (**S**). In particular, the Riemann invariants satisfy integral equations along the characteristic curves, which are well-defined for a classical solution.

Our aim is to show that also the full state is uniformly bounded in H^2 with respect to T.

For this purpose, we introduce the additional condition that $q_r(x) > 0$ for all $x \in [0, L]$. Moreover, now we assume that $C_0\varepsilon_0$ is sufficiently small so that the state constraint (5) implies that for the feasible states we have

$$q(t, x) > 0 \tag{12}$$

for all $t \in [0, T]$, $x \in [0, L]$. This is important, since the source term on the righ-hand side of (1) is in general only continuously differentiable as a function of q, since at $q = 0$ the second derivative with respect to q does not exist. Due to (12), we can also work with the second derivatives of the source term with respect to the state variables.

By our assumptions, the initial state is small in H^2 in the sense of (9).

By going to Riemann invariants $R = (R_+, R_-)$, system (1) can be written in the diagonal form

$$R_t + \Lambda(R)\, R_x = F(R)$$

with the diagonal matrix $\Lambda(R)$ that contains the eigenvalues of the system as functions of R and the function F that contains the source terms. Due to (12) we can assume that the second derivatives of F with respect to R exist for the feasible states. For a given state R, for R_x, with the notation $\hat{S} = R_x$ we obtain the semi-linear system

$$\hat{S}_t + \Lambda(R)\, \hat{S}_x = (\partial_R F(R))\hat{S} - \mathrm{diag}[(\partial_R \lambda_\pm(R))\hat{S}]\hat{S}$$

where $\partial_R F(R)$ denotes the functional matrix of F, $\lambda_\pm(R)$ denotes the eigenvalues of $\Lambda(R)$, $\partial_R \lambda_\pm(R)$ denotes the corresponding gradient and $\mathrm{diag}[(\partial_R \lambda_\pm(R))R_x]$ denotes the diagonal matrix with the corresponding entries.

For given R and R_x, for the second partial derivative R_{xx} with the notation $\bar{S} = R_{xx}$ we obtain the semi-linear system

$$\bar{S}_t + \Lambda(R)\, \bar{S}_x = (\partial_R F(R))\bar{S} + [(\partial^2_{RR} F(R))R_x]R_x$$
$$-2\,\mathrm{diag}[(\partial_R \lambda_\pm(R)) \cdot R_x]\bar{S} - \mathrm{diag}[(\partial_R \lambda_\pm(R))\bar{S}]R_x$$
$$-\mathrm{diag}\left[[(\partial^2_{RR} \lambda_\pm(R))R_x] \cdot R_x\right] R_x$$

where $[(\partial^2_{RR} F(R))R_x] = \partial_x\, \partial_R\, F(R)$ is the second derivative obtained from the chain rule and $\partial^2_{RR} \lambda_\pm(R)$ denotes the Hessian matrix of $\lambda_\pm(R)$.

Using appropriate Lyapunov functions, the semilinear systems allow to derive a priori bounds for the H^2 norm of the state in terms of the H^2 norms of the boundary data and the initial data.

Note that we know that the field of characteristics curves exists and is the same for all three systems, since we have a classical solution, For a certain fixed time $\tilde{T}$ the boundedness in H^2 of the state follows with the a priori bound described above. In order to show the uniform boundedness of the state with respect to T in H^2 we consider the problem where the roles of time and space are exchanged and u serves as 'initial' data. The a priori bound for the state that is determined by the values of

the boundary control yields the uniform boundedness of the state with respect to T.

Let us assume this in the sequel, then without loss of generality we have

$$\|p - p_r\|^2_{H^2([0,T]\times[0,L])} + \|q - q_r\|^2_{H^2([0,T]\times[0,L])} \leq C_{int}. \quad (13)$$

Then there exists $t_0 \in [0,\, T/2]$ such that

$$\|p(t_0,\cdot) - p_r\|^2_{H^2(0,L)} + \|q(t_0,\cdot) - q_r\|^2_{H^2(0,L)} \leq \frac{2\,C_{int}}{T}. \quad (14)$$

Theorem 2 implies that starting from $(\rho(t_0,\cdot),\, q(t_0,\cdot))$ the state can be controlled exactly to $(p_r,\, q_r)$ in the time T_0 with a control $u \in H^2([t_0,\, T])$ that satisfies

$$\max\{\|\hat{u} - u_r\|_{H^2(t_0,\,T)},\, \|(p - p_r, q - q_r)\|_{(H^2([t_0,\,T]\times[0,L]))^2}\}$$
$$\leq \tilde{C} \max\{\|\rho_0(t_0,\cdot) - \rho_r\|_{H^2(0,\,L)}, \|q_0(t_0,\cdot) - q_r\|_{H^2(0,\,L)}\}.$$

Now inequality (14) implies

$$\max\{\|\hat{u} - u_r\|_{H^2(t_0,\,T)},\, \|(p - p_r, q - q_r)\|_{(H^2([t_0,\,T]\times[0,L]))^2}\} \leq \frac{\sqrt{2\,C_{int}}\,\tilde{C}}{\sqrt{T}}.$$

Similar to the turnpike property with interior decay that is introduced in Gugat (2021), for $T \geq 2T_0$, this yields

$$\max\{\|\hat{u} - u_r\|_{H^2(\frac{T}{2},\,T)},\, \|(p - p_r, q - q_r)\|_{(H^2([\frac{T}{2},\,T]\times[0,L]))^2}\} \leq \frac{\sqrt{2\,C_{int}}\,\tilde{C}}{\sqrt{T}}.$$

This shows that on the second half of the time-interval, not only is the contribution of this part of the time interval uniformly bounded with respect to T but it even decays with the order $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$.

Remark. We want to mention that (12) excludes the possibility of flow reversal. Flow reversal can be very important in the operation of pipeline networks in the case of large changes in the demand scenarios. This happens in some cases regularly within the seasons of a year, but sometimes also political reasons can cause dramatic changes in the demand scenarios.

4. CONCLUSION

In this paper, we present turnpike results for a system that is governed by the isothermal Euler equations. In general this system has discontinuous solutions. However, in the operation of gas pipelines, continuously differentiable solutions are desirable and shocks are harmful in the operation of the system. We have presented an optimal control problem, where regular states are enforced by state and control constraints. We have shown that this optimal control problem has desirable turnpike properties.

ACKNOWLEDGEMENTS

This work was supported by Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Centre CRC/Transregio 154, Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks, Projects C03 and C05, Projektnummer 239904186.

REFERENCES

Banda, M.K., Herty, M., and Klar, A. (2006). Coupling conditions for gas networks governed by the isothermal Euler equations. *Netw. Heterog. Media*, 1(2), 295–314. doi:10.3934/nhm.2006.1.295.

Bastin, G. and Coron, J.M. (2016). *Stability and boundary stabilization of 1-D hyperbolic systems*, volume 88 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, [Cham]. doi:10.1007/978-3-319-32062-5. Subseries in Control.

Faulwasser, T. and Kellett, C.M. (2021). On continuous-time infinite horizon optimal control – dissipativity, stability, and transversality. *Automatica*, 134, 11. doi:10.1016/j.automatica.2021.109907. Id/No 109907.

Grüne, L., Schaller, M., and Schiela, A. (2020). Exponential sensitivity and turnpike analysis for linear quadratic optimal control of general evolution equations. *J. Differ. Equations*, 268(12), 7311–7341. doi:10.1016/j.jde.2019.11.064.

Gugat, M. and Sokolowski, J. (2022). On problems of dynamic optimal nodal control for gas networks. *Pure and Applied Functional Analysis*, x, x–x.

Gugat, M. (2021). On the turnpike property with interior decay for optimal control problems. *Mathematics of Control, Signals, and Systems*, 33, 237–258. doi:https://doi.org/10.1007/s00498-021-00280-4.

Gugat, M., Habermann, J., Hintermüller, M., and Huber, O. (2022). Constrained exact boundary controllability of a semilinear model for pipeline gas flow. *Euro. Jnl of Applied Mathematics*, 24.

Gugat, M. and Herty, M. (2022). Modeling, control, and numerics of gas networks. In E. Trelat and E. Zuazua (eds.), *Handbook of Numerical Analysis*. doi:10.1016/bs.hna.2021.12.002.

Gugat, M., Leugering, G., and Wang, K. (2017). Neumann boundary feedback stabilization for a nonlinear wave equation: a strict H^2-Lyapunov function. *Math. Control Relat. Fields*, 7(3), 419–448. doi:10.3934/mcrf.2017015.

Hafsi, Z., Mishra, M., and Elaoud, S. (2018). Hydrogen embrittlement of steel pipelines during transients. *Procedia Structural Integrity*, 13, 210–217. doi:https://doi.org/10.1016/j.prostr.2018.12.035. ECF22 - Loading and Environmental effects on Structural Integrity.

Hari, S.K.K., Sundar, K., Srinivasan, S., Zlotnik, A., and Bent, R. (2022). Operation of natural gas pipeline networks with storage under transient flow conditions. *IEEE Transactions on Control Systems Technology*, 30(2), 667–679. doi:10.1109/TCST.2021.3071316.

Li, T. (2010). *Controllability and observability for quasilinear hyperbolic systems*, volume 3 of *AIMS Series on Applied Mathematics*. American Institute of Mathematical Sciences (AIMS), Springfield, MO.

Trélat, E., Zhang, C., and Zuazua, E. (2018). Steady-state and periodic exponential turnpike property for optimal control problems in Hilbert spaces. *SIAM J. Control Optim.*, 56(2), 1222–1252. doi:10.1137/16M1097638.

Zaslavski, A.J. (2019). *Turnpike conditions in infinite dimensional optimal control*, volume 148. Cham: Springer. doi:10.1007/978-3-030-20178-4.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The Laplace transform and inconsistent initial values

Stephan Trenn

SCO @ BI(FSE), University of Groningen, Nijenborgh 9, 9747 AG Groningen, email: s.trenn@rug.nl

Abstract: Switches in electrical circuits may lead to Dirac impulses in the solution; a real world example utilizing this effect is the spark plug. Treating these Dirac impulses in a mathematically rigorous way is surprisingly challenging. This is in particular true for arguments made in the frequency domain in connection with the Laplace transform. A survey will be given on how inconsistent initials values have been treated in the past and how these approaches can be justified in view of the now available solution theory based on piecewise-smooth distributions.

Keywords: differential-algebraic equations, descriptor systems, Dirac delta, distributional solutions

1. INTRODUCTION

Modeling electrical circuits containing (ideal) switches naturally leads to a description via switched differential-algebraic equations (DAEs) of the form

$$E_{\sigma(t)}\dot{x}(t) = A_{\sigma(t)}x(t) + B_{\sigma(t)}u(t)$$

where x is the state (including algebraic variables), u is the input and σ is the switching signal (Trenn, 2012). The reason why it is in general not possible to find a more classical model in terms of ordinary differential equations (ODEs) is the fact, that the changing position of the switches changes the algebraic constraints; without including algebraic constraints in the model it would not be possible to incorporate *changing* algebraic constraints. Furthermore, the effect of *inconsistent* initial values can not be studied in a model already in the form of an ODE (because ODEs do not have inconsistent initial values). For a given switching signal σ the switched DAE can be viewed as a repeated initial value problem for the DAE

$$E\dot{x} = Ax + Bu \tag{1}$$

with initial condition $x(0) = x_0 \in \mathbb{R}^n$. As mentioned above, at a switching time the algebraic constraints may change, hence the initial value from the past may not be consistent anymore and the meaning of "$x(0) = x_0$" has to be made precise. In the context of electrical circuit this is a long standing question and can be traced back at least to Verghese et al. (1981). There have been different approaches to deal with inconsistent initial values, e.g. Sincovec et al. (1981); Cobb (1982); Opal and Vlach (1990); Rabier and Rheinboldt (1996); Reißig et al. (2002); Frasca et al. (2010), some of which will be discussed later. All have in common that jumps as well as Dirac impulses may occur in the solutions. The Dirac impulse is a distribution (a generalized function) hence one must enlarge the considered solution space to also include distributions.

This extended abstract is a revisitation of Trenn (2013) and will discuss the mathematical and conceptional difficulties arising from the notion of inconsistent initial values, in particular, when the Laplace transform is applied to the DAE (1) and arguments from the frequency domain are used. First some required notation from distribution theory is recalled, then a Laplace transform approach is presented on how to deal with inconsistent initial values and finally an alternative approach in the time domain is presented.

⋆ This work was partially supported by Vidi-grant 639.032.733

2. PRELIMINARIES ON DISTRIBUTION THEORY

The classical distribution theory by Schwartz (1957, 1959) is revised in the following. The space of *test functions* is $\mathcal{C}_0^\infty := \{ \varphi : \mathbb{R} \to \mathbb{R} \mid \varphi \in \mathcal{C}^\infty \text{ has compact support } \}$, which is equipped with a certain topology [1] . The space of distributions, denoted by $\mathbb{D}$, is then the dual of the space of test functions, i.e.

$$\mathbb{D} := \{ D : \mathcal{C}_0^\infty \to \mathbb{R} \mid D \text{ is linear and continuous } \}.$$

A large class of ordinary functions, namely locally integrable functions, can be embedded into $\mathbb{D}$ via the following injective [2] homomorphism $f \mapsto f_\mathbb{D}$ with $f_\mathbb{D}(\varphi) := \int_\mathbb{R} f\varphi$.

The main feature of distributions is the ability to take derivatives for any distribution $D \in \mathbb{D}$ via $D'(\varphi) := -D(\varphi')$, which is consistent with the classical derivative, i.e. if f is differentiable, then $(f_\mathbb{D})' = (f')_\mathbb{D}$. In particular, the Heaviside unit step $\mathbb{1}_{[0,\infty)}$ has a distributional derivative which can easily be calculated to be

$$(\mathbb{1}_{[0,\infty)_\mathbb{D}})'(\varphi) = \varphi(0) =: \delta(\varphi),$$

hence it results in the well known Dirac impulse δ (at $t = 0$). In general, the Dirac impulse δ_t at time $t \in \mathbb{R}$ is given by $\delta_t(\varphi) := \varphi(t)$. Furthermore, if g is a piecewise differentiable function with one jump at $t = t_J$, then

$$(g_\mathbb{D})' = (g')_\mathbb{D} + (g(t_J^+) - g(t_J^-))\delta_{t_J}, \tag{2}$$

where g' is the derivative of g on $\mathbb{R} \setminus \{0\}$.

[1] The topology is such that a sequence $(\varphi_k)_{k\in\mathbb{N}}$ of test functions converges to zero if, and only if, 1) the supports of all φ_k are contained within one common compact set $K \subseteq \mathbb{R}$ and 2) for all $i \in \mathbb{N}$, $\varphi_k^{(i)}$ converges uniformly to zero as $k \to \infty$

[2] Two locally integrable functions which only differ on a set of measure zero are identified with each other.

Now it is no problem to consider the DAE (1) (without the initial condition) in a distributional solution space; instead of x and u being vectors of functions they are now vectors of distributions, i.e. $x \in \mathbb{D}^n$ and $f \in \mathbb{D}^m$ where $n \times n$ and $n \times m$ are the size of the matrices E, A and B. The definition of the matrix vector product remains unchanged [3] so that (1) reads as m equations in $\mathbb{D}$.

Considering distributional solutions, however, does *not* help to treat inconsistent initial value; au contraire, distributions cannot be evaluated at a certain time because they are not functions of time, so writing $x(0) = x_0$ makes no sense. Even when assuming that a pointwise evaluation is well defined for certain distributions, the DAE (1) will still not exhibit (distributional) solution with arbitrary initial values. This is easily seen when considering, e.g., the DAE (1) with $(E, A, B) = (0, I, 0)$, which simply reads as $0 = x$.

So what does it then mean to speak of a solution of (1) with inconsistent initial value? The motivation for inconsistent initial values is the situation that the system descriptions gets active at the initial time $t = 0$ and before that the system was governed by different (maybe unknown) rules. This viewpoint was already expressed by Doetsch (1974) in the context of distributional solutions for ODEs:

> The concept of "initial value" in the physical science can be understood only when the past, that is the interval $t < 0$, has been included in our considerations. This occurs naturally for distributions which, without exception, are defined on the entire t-axis.

So mathematically, there is some given past trajectory x^0 for x up to the initial time and the DAE (1) only holds on the interval $[0, \infty)$. This means that a solution of the following *initial trajectory problem* (ITP) is sought:

$$\begin{aligned} x_{(-\infty,0)} &= x^0_{(-\infty,0)} \\ (E\dot{x})_{[0,\infty)} &= (Ax + Bu)_{[0,\infty)}, \end{aligned} \tag{3}$$

where $x^0 \in \mathbb{D}^n$ is an arbitrary past trajectory and D_I for some interval $I \subseteq \mathbb{R}$ and $D \in \mathbb{D}$ denotes a distributional restriction generalizing the restrictions of functions given by$f_I(t) = f(t)$ for $t \in I$ and $f(t) = 0$ otherwise.

A fundamental problem is the fact (Trenn, 2021) that such a distributional restriction does not exist!

This problem was resolved especially in older publication (Campbell, 1980, 1982; Verghese et al., 1981) by ignoring it and/or by arguing with the Laplace transform (see the next section). Cobb (1984) seems to be the first to be aware of this problem and he resolved it by introducing the space of piecewise-continuous distributions; Geerts (1993b,a) was the first to use the space of impulsive-smooth distributions (introduced by Hautus and Silverman (1983)) as a solution space for DAEs. Seemingly unaware of these two approaches, Tolsa and Salichs (1993) developed a distributional solution framework which can be seen as a mixture between the approaches of Cobb and Geerts. The more comprehensive space of piecewise-smooth distributions was later introduced (Trenn, 2009) to combine the advantages of the piecewise-continuous and impulsive-smooth distributional solution spaces. Further details are discussed in Section 4.

Cobb (1982) also presented another approach by justifying the impulsive response due to inconsistent initial values via his notion of *limiting solutions*. The idea is to replace the singular matrix E in (1) by a "disturbed" version E_ε which is invertible for all $\varepsilon > 0$ and $E_\varepsilon \to E$ as $\varepsilon \to 0$. If the solutions of the corresponding initial value ODE problem $\dot{x} = E_\varepsilon^{-1}Ax$, $x(0) = x_0$ converges to a distribution, then Cobb calls this the limiting solution. He is then able to show that the limiting solution is unique and equal to the one obtained via the Laplace-transform approach. Campbell (1982) extends this result also to the inhomogeneous case.

3. LAPLACE TRANSFORM APPROACHES

Especially in the signal theory community it is common to study systems like (1) in the so called *frequency domain* (in contrast to the *time domain*). The transformation between time and frequency domain is given by the *Laplace transform* defined via the Laplace integral:

$$\hat{g}(s) := \int_0^\infty e^{-st} g(t) \ \mathrm{d}t \tag{4}$$

for some function g and $s \in \mathbb{C}$. Note that in general the Laplace integral is not well defined for all $s \in \mathbb{C}$ and a suitable domain for $\hat{g}$ must be chosen (Doetsch, 1974). If a suitable domain exists, then $\hat{g} = \mathcal{L}\{g\}$ is called the *Laplace transform* of g and, in general, $\mathcal{L}\{\cdot\}$ denotes the Laplace transform operator. Again note that it is not specified at this point which class of functions have a Laplace transform and which class of functions are obtained as the image of $\mathcal{L}\{\cdot\}$. The main feature of the Laplace transform is the following property, where g is a differentiable function for which g and g' have Laplace transforms,

$$\mathcal{L}\{g'\}(s) = s\mathcal{L}\{g\}(s) - g(0), \tag{5}$$

which is a direct consequence of the definition of the Laplace integral invoking partial differentiation. If g is not continuous at $t = 0$ but $g(0^+)$ exists and g' denotes the derivative of g on $\mathbb{R} \setminus \{0\}$, then (5) still holds in a slightly altered form:

$$\mathcal{L}\{g'\}(s) = s\mathcal{L}\{g\}(s) - g(0^+). \tag{6}$$

In particular, the Laplace transform does not take into account at all how g behaved for $t < 0$ which is a trivial consequence of the definition of the Laplace integral. This observation will play an important role when studying inconsistent initial values.

Taking into account the linearity of the Laplace transform the DAE (1) is transformed into

$$sE\hat{x}(s) = A\hat{x}(s) + B\hat{u}(s) + Ex(0^+) \tag{7}$$

If the matrix pair (E, A) is regular [4] and $x(0^+) = 0$, the latter can be solved easily algebraically:

$$\hat{x}(s) = (sE - A)^{-1}B\hat{u}(s) =: G(s)\hat{u}(s), \tag{8}$$

where $G(s)$ is a matrix over the field of rational functions and is usually called transfer function.

[3] Some authors (Rabier and Rheinboldt, 2002; Kunkel and Mehrmann, 2006) use a different definition for the matrix vector product which is due to the different viewpoint of a distributional vector x as a map from $(\mathcal{C}_0^\infty)^n$ to $\mathbb{R}$ instead of a map from $\mathcal{C}_0^\infty$ to $\mathbb{R}^n$. The latter seems the more natural approach in view of applying it to (1), but it seems that both approaches are equivalent at least with respect to the solution theory of DAEs.

[4] (E, A) is called regular, if $\det(sE - A)$ is not identically zero.

A first systematic treatment of descriptor systems in the frequency domain was carried out by Rosenbrock (1970). He, however, only considered zero initial values and the input-output behavior. In particular, he was not concerned with a solution theory for general DAEs (1) with possible inconsistent values. Furthermore, he restricted attention to inputs which are exponentially bounded (guaranteeing existence of the Laplace transform), hence formally his framework could not deal with arbitrary (sufficiently smooth) inputs.

The definition of the Laplace transform can be extended to be well defined for certain distributions as well (Doetsch, 1974), therefore consider the following class of distributions:

$$\mathbb{D}_{\geq 0,k} := \left\{ D=(g_{\mathbb{D}})^{(k)} \;\middle|\; \begin{array}{l}\text{where } g:\mathbb{R}\to\mathbb{R} \text{ is continuous}\\ \text{and } g(t)=0 \text{ on } (-\infty,0)\end{array} \right\}.$$

For $D \in \mathbb{D}_{\geq 0,k}$ with $D = (g_{\mathbb{D}})^{(k)}$ the (distributional) Laplace transform is now given by

$$\mathcal{L}_{\mathbb{D}}\{D\}(s) := s^k \mathcal{L}\{g\}(s)$$

on a suitable domain in $\mathbb{C}$. Note that $\delta \in \mathbb{D}_{\geq 0,2}$ and it is easily seen that

$$\mathcal{L}_{\mathbb{D}}\{\delta\} = 1. \tag{9}$$

Furthermore, for every locally integrable function g for which $\mathcal{L}\{g\}$ is defined on a suitable domain it holds

$$\mathcal{L}_{\mathbb{D}}\{g_{\mathbb{D}}\} = s\mathcal{L}\left\{\int_0^{\cdot} g\right\} = s\frac{1}{s}\mathcal{L}\{g\} = \mathcal{L}\{g\}, \tag{10}$$

i.e. the distributional Laplace transform coincides with the classical Laplace transform defined by (4).

A direct consequence of the definition of $\mathcal{L}_{\mathbb{D}}$ is the following derivative rule for all $D \in \bigcup_k \mathbb{D}_{\geq 0,k}$:

$$\mathcal{L}_{\mathbb{D}}\{D'\}(s) = s\mathcal{L}_{\mathbb{D}}\{D\} \tag{11}$$

which seems to be in contrast to the derivative rule (6), because *no initial value occurs*. The latter can actually not be expected because general distributions do not have a well defined function evaluation at a certain time t. However, the derivative rule (11) is consistent with (6); to see this let g be a function being zero on $(-\infty, 0)$, differentiable on $(0,\infty)$ with well defined value $g(0^+)$. Denote with g' the (classical) derivative of g on $\mathbb{R}\setminus\{0\}$, then (invoking linearity of $\mathcal{L}_{\mathbb{D}}$)

$$\begin{aligned}\mathcal{L}_{\mathbb{D}}\{(g_{\mathbb{D}})'\} &\overset{(2)}{=} \mathcal{L}_{\mathbb{D}}\{(g')_{\mathbb{D}} + g(0^+)\delta\}\\ &= \mathcal{L}_{\mathbb{D}}\{(g')_{\mathbb{D}}\} + g(0^+)\mathcal{L}_{\mathbb{D}}\{\delta\} \overset{(9),(10)}{=} \mathcal{L}\{g'\} + g(0^+),\end{aligned}$$

which shows equivalence of (11) and (6). The key observation is that the distributional derivative takes into account the jump at $t = 0$ whereas the classical derivates ignores it, i.e. in the above context

$$(g_{\mathbb{D}})' \neq (g')_{\mathbb{D}}.$$

As it is common to identify g with $g_{\mathbb{D}}$ (even in Doetsch (1974)), the above distinction is difficult to grasp, in particular for inexperienced readers. As this problem plays an important role when dealing with inconsistent initial values, it is not surprising that researchers from the DAE community who are simply using the Laplace transform as a tool, struggle with the treatment of inconsistent initial values, c.f. Lundberg et al. (2007).

Revisiting the treatment of the DAE (1) in the frequency domain one has now to decide wether to use the usual Laplace transform resulting in (7) or the distributional Laplace transform resulting in

$$sE\hat{x}(s) = A\hat{x}(s) + B\hat{u}(s), \tag{12}$$

where the initial value $x(0^+)$ does not occur anymore. In particular, if $u = 0$ the only solution of (12) is $\hat{x}(s) = 0$, which implies $x = 0$. Altogether, the following dilemma occurs:

Dilemma. Consider the regular DAE (1) with zero input but non-zero initial value, then the following conflicting observations can be made:
• An adhoc analysis calls for *distributional solutions* in response to inconsistent initial values. For consistent initial value there exist classical (nonzero) solutions.
• Using the *distributional* Laplace transform to analyze the (distributional) solutions of (1) reveals that the *only* solution is the trivial one. In particular, no initial values (neither inconsistent nor consistent ones) are taken into account at all.

This problem was already observed by Doetsch (1974) and is based on the definition of the distributional Laplace transform which is only defined for distributions vanishing on $(-\infty, 0)$. The following "solution" to this Dilemma was suggested (Doetsch, 1974, p. 129): Define for $D \in \bigcup_k \mathbb{D}_{\geq 0,k}$ the "past-aware" derivative operator $\frac{\mathrm{d}_-}{\mathrm{d}t}$:

$$\tfrac{\mathrm{d}_-}{\mathrm{d}t} D := D' - d_0^- \delta \tag{13}$$

where $d_0^- \in \mathbb{R}$ is interpreted as a "virtual" initial value for $D(0^-)$. Note however, that by definition $D(0^-) = 0$ for every $D \in \bigcup_k \mathbb{D}_{\geq 0,k}$, hence at this stage it is not clear why this definition makes sense. This problem was also pointed out by Cobb (1982).

Using now the past-aware derivative in the distributional formulation of (1) one obtains:

$$Ex' = Ax + Bu + Ex_0^-\delta \tag{14}$$

where $x_0^- \in \mathbb{R}^n$ is the virtual (possible inconsistent) initial value for $x(0^-)$ and solutions are sought in the space $\left(\bigcup_k \mathbb{D}_{\geq 0,k}\right)^n$, i.e. x is assumed to be zero on $(-\infty, 0)$. Applying the distributional Laplace transform to (14) yields

$$sE\hat{x}(s) = A\hat{x}(s) + B\hat{u}(s) + Ex_0^- \tag{15}$$

In contrast to (7), x_0^- is not the initial value for $x(0^+)$ but is the virtual initial value for $x(0^-)$. If the matrix pair (E, A) is regular, the solution of (15) can now be obtained via $\hat{x}(s) = (sE-A)^{-1}(B\hat{u}(s)+Ex_0^-)$ and using the inverse Laplace transform; there are however the following major drawbacks:
(i) Within the frequency domain it is not possible to motivate the incorporation of the (inconsistent) initial values as in (14); in fact, Doetsch (1974) who seems to have introduced this notion needs to argue with the help of the distributional derivative and (13) within the time domain!
(ii) The Laplace transform ignores everything what was in the past, i.e. on the interval $(-\infty, 0)$; this is true for the classical Laplace transform (by definition of the Laplace integral) as well as for the distributional Laplace transform (by only considering distributions which vanish for $t < 0$). Hence the natural viewpoint of an initial trajectory problem (3) as also informally advocated by Doetsch is not possible to treat with the Laplace transform approach.
(iii) Making statements about existence and uniqueness

of solution with the help of the frequency domain heavily depends on an isomorphism between the time-domain and the frequency domain; there are, however, only a few special isomorphisms between certain special subspaces of the frequency and time domain, no general isomorphismus is available.

4. PIECEWISE-SMOOTH DISTRIBUTIONS

In order to rigorously analyse switched DAEs it was suggested in Trenn (2009) to use as an underlying solution space the space of piecewise-smooth distributions

$$\mathbb{D}_{\mathrm{pw}\mathcal{C}^\infty} := \left\{ D = f_{\mathbb{D}} + \sum_{t\in T} D_t \;\middle|\; \begin{array}{l} f \in \mathcal{C}^\infty_{\mathrm{pw}}, T \subseteq \mathbb{R} \text{ locally} \\ \text{finite, } \forall t \in T: \\ D_t \in \operatorname{span}\{\delta_t, \delta_t', \delta_t'', \ldots\} \end{array} \right\},$$

where $\mathcal{C}^\infty_{\mathrm{pw}}$ is the space of piecewise-smooth functions (with locally finitely many discontinuities). This space is closed under differentiation and therefore removes one shortcoming of Cobb's space of piecewise-continuous distributions and generalized the space of impulsive-smooth distributions, which only considers Dirac impulses at $t = 0$ [5].

A key result for the ITP (3) is then the following equivalence:

Theorem 1. (cf. Thm. 5.3 in Trenn (2013)). Consider the ITP (3) within the piecewise-smooth distributional solution framework with fixed initial trajectory $x^0 \in \mathbb{D}^n_{\mathrm{pw}\mathcal{C}^\infty}$ and inhomogeneity $u \in \mathbb{D}^m_{\mathrm{pw}\mathcal{C}^\infty}$. Then $x \in \mathbb{D}^n_{\mathrm{pw}\mathcal{C}^\infty}$ solves the ITP (3) if, and only if, $z := x - x^0_{(-\infty,0)} = x_{[0,\infty)}$ solves

$$\begin{aligned} z_{(-\infty,0)} &= 0 \\ (E\dot{z})_{[0,\infty)} &= (Az + Bu)_{[0,\infty)} + Ex^0(0-)\delta. \end{aligned} \tag{16}$$

Corollary 2. Consider a (possible inconsistent) initial value $x_0 \in \mathbb{R}^n$ for the regular DAE (1). Then for any trajectory $x^0 \in \mathbb{D}^n_{\mathrm{pw}\mathcal{C}^\infty}$ with $x^0(0^-) = x_0$ and any input u, the solution x of the ITP (3) restricted to $[0,\infty)$ equals the solution obtained via the Laplace transform approach (15) (under the assumption $(x,u)_{[0,\infty)} \in \left(\bigcup_k \mathbb{D}_{\geq 0,k}\right)^{n\times m}$).

5. CONCLUSION

Inconsistent initial values cannot be treated in a meaningful way when studying DAEs in the frequency domain. However, arguments in the time-domain based on piecewise-smooth distribution justify why adding the term Ex_0 to the right-hand side of the distributionally Laplace transformed DAE indeed results in a meaningful solution to the inconsistent initial value problem

REFERENCES

Campbell, S.L. (1980). *Singular Systems of Differential Equations I.* Pitman, New York.

Campbell, S.L. (1982). *Singular Systems of Differential Equations II.* Pitman, New York.

Cobb, J.D. (1982). On the solution of linear differential equations with singular coefficients. *J. Diff. Eqns.*, 46, 310–323. doi:10.1016/0022-0396(82)90097-3.

Cobb, J.D. (1984). Controllability, observability and duality in singular systems. *IEEE Trans. Autom. Control*, 29, 1076–1082. doi:10.1109/TAC.1984.1103451.

Doetsch, G. (1974). *Introduction to the Theory and Application of the Laplace Transformation.* Springer-Verlag, Berlin.

Frasca, R., Çamlıbel, M.K., Goknar, I.C., Iannelli, L., and Vasca, F. (2010). Linear passive networks with ideal switches: Consistent initial conditions and state discontinuities. *IEEE Trans. Circuits Syst., I: Fundam. Theory Appl.*, 57(12), 3138–3151.

Geerts, A.H.W.T. (1993a). Invariant subspaces and invertibility properties for singular systems: the general case. *Linear Algebra Appl.*, 183, 61–88. doi:10.1016/0024-3795(93)90424-M.

Geerts, A.H.W.T. (1993b). Solvability conditions, consistency and weak consistency for linear differential-algebraic equations and time-invariant linear systems: The general case. *Linear Algebra Appl.*, 181, 111–130. doi:10.1016/0024-3795(93)90027-L.

Geerts, A.H.W.T. and Schumacher, J.M.H. (1996). Impulsive-smooth behavior in multimode systems. Part I: State-space and polynomial representations. *Automatica*, 32(5), 747–758.

Hautus, M.L.J. and Silverman, L.M. (1983). System structure and singular control. *Linear Algebra Appl.*, 50, 369–402.

Kunkel, P. and Mehrmann, V. (2006). *Differential-Algebraic Equations. Analysis and Numerical Solution.* EMS Publishing House, Zürich, Switzerland. doi:10.4171/017.

Lundberg, K.H., Miller, H.R., and Trumper, D.L. (2007). Initial conditions, generalized functions, and the Laplace transform. *IEEE Control Systems Magazine*, 27(1), 22–35. doi:10.1109/MCS.2007.284506.

Opal, A. and Vlach, J. (1990). Consistent initial conditions of linear switched networks. *IEEE Trans. Circuits Syst.*, 37(3), 364–372.

Rabier, P.J. and Rheinboldt, W.C. (1996). Time-dependent linear DAEs with discontinuous inputs. *Linear Algebra Appl.*, 247, 1–29.

Rabier, P.J. and Rheinboldt, W.C. (2002). Theoretical and numerical analysis of differential-algebraic equations. In P.G. Ciarlet and J.L. Lions (eds.), *Handbook of Numerical Analysis*, volume VIII, 183–537. Elsevier Science, Amsterdam, The Netherlands.

Reißig, G., Boche, H., and Barton, P.I. (2002). On inconsistent initial conditions for linear time-invariant differential-algebraic equations. *IEEE Trans. Circuits Syst., I: Fundam. Theory Appl.*, 49(11), 1646–1648.

Rosenbrock, H.H. (1970). *State Space and Multivariable Theory.* John Wiley and Sons Inc., New York, NY.

Schwartz, L. (1957, 1959). *Théorie des Distributions.* Hermann, Paris.

Sincovec, R.F., Erisman, A.M., Yip, E.L., and Epton, M.A. (1981). Analysis of descriptor systems using numerical algorithms. *IEEE Trans. Autom. Control*, 26, 139–147.

Tolsa, J. and Salichs, M. (1993). Analysis of linear networks with inconsistent initial conditions. *IEEE Trans. Circuits Syst.*, 40(12), 885 – 894. doi:10.1109/81.269029.

Trenn, S. (2009). *Distributional differential algebraic equations.* Ph.D. thesis, Institut für Mathematik, Technische Universität Ilmenau, Universitätsverlag Ilmenau, Germany. URL `http://www.db-thueringen.de/servlets/DocumentServlet?id=13581`.

Trenn, S. (2012). Switched differential algebraic equations. In F. Vasca and L. Iannelli (eds.), *Dynamics and Control of Switched Electronic Systems - Advanced Perspectives for Modeling, Simulation and Control of Power Converters*, chapter 6, 189–216. Springer-Verlag, London. doi:10.1007/978-1-4471-2885-4_6.

Trenn, S. (2013). Solution concepts for linear DAEs: a survey. In A. Ilchmann and T. Reis (eds.), *Surveys in Differential-Algebraic Equations I*, Differential-Algebraic Equations Forum, 137–172. Springer-Verlag, Berlin-Heidelberg. doi:10.1007/978-3-642-34928-7_4.

Trenn, S. (2021). Distributional restriction impossible to define. *Examples and Counterexamples*, 1(100023), 1–4. doi:10.1016/j.exco.2021.100023. Open access.

Verghese, G.C., Levy, B.C., and Kailath, T. (1981). A generalized state-space for singular systems. *IEEE Trans. Autom. Control*, 26(4), 811–831.

[5] Rabier and Rheinboldt (1996) seem to be aware of this restriction and they introduce the space $\mathcal{C}_{\mathrm{imp}}(\mathbb{R}\setminus\mathcal{S})$, where $\mathcal{S} = \{\, t_i \in \mathbb{R} \mid i \in \mathbb{Z} \,\}$ is a strictly ordered set with $t_i \to \pm\infty$ as $i \to \pm\infty$ and $D \in \mathcal{C}_{\mathrm{imp}}(\mathbb{R}\setminus\mathcal{S})$ is such that $D_{(t_i,t_{i+1})}$ is induced by the corresponding restriction of a smooth function. A similar idea is proposed in Geerts and Schumacher (1996), however in both cases the resulting distributional space is not studied in detail.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Matrix Pontryagin principle approach to controllability metrics maximization under sparsity constraints ⋆

Tomofumi Ohtsuka (Kyoto University), *
Takuya Ikeda (The University of Kitakyushu), **
Kenji Kashima (Kyoto University) *

* *Graduate School of Informatics, Kyoto University, Kyoto, Japan.*
(e-mail: otsuka.tomofumi@bode.amp.i.kyoto-u.ac.jp, kk@i.kyoto-u.ac.jp)
** *Faculty of Environmental Engineering, The University of Kitakyushu,*
Fukuoka, Japan. (e-mail: t-ikeda@kitakyu-u.ac.jp)

Abstract: Controllability maximization problem under sparsity constraints is a node selection problem that selects inputs that are effective for control in order to minimize the energy to control for desired state. In this paper we discuss the equivalence between the sparsity constrained controllability metrics maximization problems and their convex relaxation. The proof is based on the matrix-valued Pontryagin maximum principle applied to the controllability Lyapunov differential equation.

Keywords: Sparse optimal control, node selection problem, controllability maximization

1. INTRODUCTION

Sparse optimal problems have attracted a lot of attention in the field of optimal control. Such an approach is useful to find a small number of essential information that is closely related to the control performance of interest, and it is applied widely, for example, Ikeda et al. (2021). This paper investigates the application of sparse optimization to controllability maximization problem, one of the control node selection problems. The problem is known as the optimization problem minimizing the energy to control for the desired state; see also Olshevsky (2014) and Pasqualetti et al. (2014) for other related metrics.

These problems are generally formulated as maximization of some metric of the controllability Gramian with L^0/l^0 constraints, but it is known that the problems include combinatorial structures. To circumvent this, relaxed problems, where the L^0/l^0 norms are replaced by the L^1/l^1 norms, are considered for its computational tractability. Then, the problem is how to prove the equivalence between the main problem and its relaxation. The paper Ikeda and Kashima (2018) proved the equivalence when the trace of controllability Gramian is adopted as the metric, but its usefulness as a metric is questionable since the designed Gramian may include the zero eigenvalue, so the trace metric does not automatically ensure the controllability. The paper Ikeda and Kashima (2022) considered the minimum eigenvalue and the determinant of the controllability Gramian which is useful as metrics, but it avoided the proof of equivalence because of the difficulty and treated approximation problems that are easy to prove the equivalence. In view of this, this paper newly proposes a method to prove the equivalence for general metrics of controllability. Specifically, we adopted the controllability Lyapunov differential equation. The controllability Lyapunov differential equation is a matrix-valued differential equation whose solution is the controllability Gramian. By considering the optimal control problem for this Lyapunov differential equation, we can strictly treat useful metrics that are related to the controllability Gramian.

The remainder of this paper is organized as follows. Section 2 provides mathematical preliminaries. Section 3 formulates our node scheduling problem using controllability Lyapunov differential equation, and gives a sufficient condition for the main problem to boil down to the relaxation problem. Section 4 offers concluding remarks.

Notation

For any $A, B \in \mathbb{R}^{n\times m}$, we denote the Frobenius norm of A by $\|A\| \triangleq \left(\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2\right)^{1/2}$, and the inner product of A and B by $(A, B) \triangleq \left(\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}B_{i,j}\right)$. Let C be a closed subset of $\mathbb{R}^{n\times m}$ and $A \in C$. A matrix $\Delta \in \mathbb{R}^{n\times m}$ is a *proximal normal* to the set C at the point A if and only if there exists a constant $\sigma \geq 0$ such that $(\Delta, B - A) \leq \sigma\|B - A\|^2$ for all $B \in C$. The *proximal normal cone* to C at A is defined as the set of all such Δ, which is denoted by $N_C^P(A)$. We denote the *limiting normal cone* to C at A by $N_C^L(A)$, i.e., $N_C^L(A) \triangleq \{\Delta = \lim_{i\to\infty}\Delta_i : \Delta_i \in N_C^P(A_i), A_i \to A, A_i \in C\}$. For other notations, see (Ikeda and Kashima, 2022, Section II).

2. PRELIMINARY

Let us consider the following continuous-time linear system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BV(t)u(t), \quad t \in [0, T], \\ V(t) &= \mathrm{diag}(v(t)), \quad v(t) \in \{0, 1\}^p, \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector consisting of n nodes, where $x_i(t)$ is the state of the i-th node at time t; $u(\cdot) \in \mathbb{R}^m$ is the exogenous control input that influences the network dynamics. Then the controllability Gramian for the system is defined by

⋆ This work was supported in part by JSPS KAKENHI under Grant Number JP18H01461 and JP21H04875.

$$G_c = \int_0^T e^{A(T-\tau)} BV(\tau)V(\tau)^\top B^\top e^{A^\top (T-\tau)} d\tau. \tag{2}$$

We next show why the controllability Gramian is used as the metric of the ease of control. We here recall the minimum-energy control problem:

$$\begin{aligned} \min_u \quad & \int_0^T \|u(t)\|^2 dt \\ \text{s.t.} \quad & \dot{x}(t) = Ax(t) + BV(t)u(t), \\ & x(0) = 0, \quad x(T) = x_f. \end{aligned} \tag{3}$$

The minimum control energy is then given by $x_f^\top G_c^{-1} x_f$ (Verriest and Kailath (1983)). Based on this, recent works have been considered to make G_c as large as possible. In this paper we design $BV(t)$ in order to maximize some metric of the controllability Gramian. As the constraints, we introduce L^0 and l^0 constraints on $v(t)$ to take account of the upper bound of the total time length of node activation and the number of activated nodes at each time. We consider the following optimal problem that maximizes some metric of G_c under sparsity constraints:

$$\begin{aligned} \max_v \quad & J(v) = K(G_c) \\ \text{s.t.} \quad & v(t) \in \{0,1\}^p \quad {}^\forall t \in [0,T], \\ & \|v_j\|_{L^0} \le \alpha_j \quad {}^\forall j \in \{1,2,\dots,p\}, \\ & \|v(t)\|_{l^0} \le \beta \quad {}^\forall t \in [0,T], \end{aligned} \tag{4}$$

where $K(G_c)$ is a metric of the controllability Gramian, and $\alpha_j > 0$ and $\beta > 0$ is constant.

Since the maximization problem in (4) is a combinatorial optimization problem, we consider the following relaxation problem:

$$\begin{aligned} \max_v \quad & J(v) = K(G_c) \\ \text{s.t.} \quad & v(t) \in [0,1]^p \quad {}^\forall t \in [0,T], \\ & \|v_j\|_{L^1} \le \alpha_j \quad {}^\forall j \in \{1,2,\dots,p\}, \\ & \|v(t)\|_{l^1} \le \beta \quad {}^\forall t \in [0,T]. \end{aligned} \tag{5}$$

This problem is easier to treat than the main problem (especially if K is concave, problem (5) is a convex optimization problem). We, however, have to consider the equivalence between the main problem and the corresponding relaxation problem. Ikeda and Kashima (2022) formulated alternative approximation problem instead of proving the equivalence. Then this paper proves the equivalence between the main problem and the relaxed one by using the controllability Lyapunov differential equation.

3. PROPOSED METHOD

In this section, we formulate a controllability Lyapunov differential equation which holds the controllability Gramian as a solution, and then formulate an optimization problem for a system in which the state space representation is given by the derived differential equation. We provide an equivalence theorem between the main problem and the corresponding relaxation problem.

3.1 Problem formulation and relaxation

Controllability Lyapunov differential equation is given as follows:

$$\begin{aligned} & \dot{G}_c(t) = AG_c(t) + G_c(t)A^\top + BV(t)V(t)^\top B^\top, \\ & G_c(0) = O_{n\times n}. \end{aligned} \tag{6}$$

Then the controllability Gramian G_c defined by (2) corresponds to the solution $G_c(T)$ of (6) at $t = T$. Here we consider the following optimal control problem.

Problem 1 (Main problem)**.**

$$\begin{aligned} \max_v \quad & J(v) = K(G_c(T)) \\ \text{s.t.} \quad & \dot{G}_c(t) = AG_c(t) + G_c(t)A^\top + BV(t)B^\top, \\ & G_c(0) = O_{n\times n}, \\ & v(t) \in \{0,1\}^p \quad {}^\forall t \in [0,T], \\ & \|v_j\|_{L^0} \le \alpha_j \quad {}^\forall j \in \{1,2,\dots,p\}, \\ & \|v(t)\|_{l^0} \le \beta \quad {}^\forall t \in [0,T]. \end{aligned} \tag{7}$$

Note that $V(\cdot)V(\cdot)^\top = V(\cdot)$ since $v(\cdot) \in \{0,1\}^p$, so we rewrite the controllability Lyapunov differential equation. Problem 1 is a combinatorial optimization problem, so we consider the following relaxed problem, where the L^0/l^0 norms are replaced by the L^1/l^1 norms, respectively.

Problem 2 (Relaxed problem)**.**

$$\begin{aligned} \max_v \quad & J(v) = K(G_c(T)) \\ \text{s.t.} \quad & \dot{G}_c(t) = AG_c(t) + G_c(t)A^\top + BV(t)B^\top, \\ & G_c(0) = O_{n\times n}, \\ & v(t) \in [0,1]^p \quad {}^\forall t \in [0,T], \\ & \|v_j\|_{L^1} \le \alpha_j \quad {}^\forall j \in \{1,2,\dots,p\}, \\ & \|v(t)\|_{l^1} \le \beta \quad {}^\forall t \in [0,T]. \end{aligned} \tag{8}$$

In what follows, we suppose that K is continuously differentiable.

3.2 discreteness and equivalence

We define the set of feasible solutions of Problem 1 and Problem 2 by $\mathcal{V}_0$ and $\mathcal{V}_1$, i.e.,

$$\begin{aligned} \mathcal{V}_0 \triangleq \{v : & v(t) \in \{0,1\}^p \quad {}^\forall t, \quad \|v_j\|_{L^0} \le \alpha_j \quad {}^\forall j, \\ & \|v(t)\|_{l^0} \le \beta \quad {}^\forall t\}, \\ \mathcal{V}_1 \triangleq \{v : & v(t) \in [0,1]^p \quad {}^\forall t, \quad \|v_j\|_{L^1} \le \alpha_j \quad {}^\forall j, \\ & \|v(t)\|_{l^1} \le \beta \quad {}^\forall t\}. \end{aligned}$$

Note that $\mathcal{V}_0 \subset \mathcal{V}_1$, since $\|v_j\|_{L^1} = \|v_j\|_{L^0}$ for all j and $\|v(t)\|_{l^1} = \|v(t)\|_{l^0}$ on $[0,T]$ for any measurable function v with $v(t) \in \{0,1\}^p$ on $[0,T]$. The inclusion is proper in general, since the L^1/l^1 constraints do not automatically guarantee the L^0/l^0 constraints and some functions in $\mathcal{V}_1$ are not obviously binary. Then, we first show the discreteness of solutions of Problem 2, which guarantees that the optimal solutions of Problem 2 belongs to the set $\mathcal{V}_0$. For this purpose, we prepare lemmas.

Lemma 1 (Matrix Pontryagin principle)**.** *Let us consider the following optimization problem*

$$\begin{aligned} \min_U \quad & J = L_f(X(T)) \\ \text{s.t.} \quad & \dot{X}(t) = F(X(t), U(t)), \\ & X(0) = X_0, \quad X(T) \in E, \quad U(t) \in \Omega, \end{aligned} \tag{9}$$

where L_f is continuously differentiable, F is continuous, $D_X F(X(t), U(t))$ is continuous with respect to t, X, U, $X(t) \in \mathbb{R}^{n\times m}$, $U(t) \in \mathbb{R}^{p\times q}$, $X_0 \in \mathbb{R}^{n\times m}$, $T > 0$, $E \subset \mathbb{R}^{n\times m}$, and $\Omega \subset \mathbb{R}^{p\times q}$. Note that $(L_f, F, X_0, T, E, \Omega)$ is given. We define Hamiltonian function $H : \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times q} \to \mathbb{R}$ associated to problem (9) by

$$H(X(t), P(t), U(t)) = \mathrm{Tr}\left(P(t)^\top F(X(t), U(t))\right). \tag{10}$$

Let the process $(X^*(t), V^*(t))$ be a local minimizer for the problem (9). Then there exists a matrix $P : [0,T] \to \mathbb{R}^{n\times m}$, and a scalar η equal to 0 or 1 satisfying the following conditions:

- *the nontriviality condition:*
$$(\eta, P(t)) \neq 0 \quad {}^\forall t \in [0,T], \tag{11}$$
- *the transversality condition:*
$$-P(T) \in \eta \nabla L_f(X^*(T)) + N_E^L(X^*(T)), \tag{12}$$
- *the adjoint equation for almost every $t \in [0,T]$:*
$$-\dot{P}(t) = D_X H(X^*(t), P(t), U^*(t)), \tag{13}$$
- *the maximum condition for almost every $t \in [0,T]$:*
$$H(X^*(t), P(t), U^*(t)) = \sup_{U\in\Omega} H(X^*(t), P(t), U). \tag{14}$$

Proof. We define a mapping $\psi_{nm} : \mathbb{R}^{n\times m} \to \mathbb{R}^{nm}$ by
$$\psi_{nm}(X) = \left[X_1^\top, \ldots, X_m^\top\right]^\top, \tag{15}$$
where $X_i \in \mathbb{R}^n$ denotes the ith column of a matrix $X \in \mathbb{R}^{n\times m}$. From Athans (1967), ψ_{nm} is a regular linear mapping (hence ψ_{nm}^{-1} exists), and preserves the inner product. Then problem (9) is equivalent to
$$\begin{aligned} \min_u \quad & J = l_f(x(T)) \\ \text{s.t.} \quad & \dot{x}(t) = f(x(t), u(t)), \\ & x(0) = x_0 \quad x(T) \in e, \quad u(t) \in \omega, \end{aligned} \tag{16}$$
where $x = \psi_{nm}(X), u = \psi_{pq}(U)$, and l_f, f, x_0, e, ω corresponds to L_f, F, X_0, E, Ω respectively. We define the Hamiltonian function $h : \mathbb{R}^{nm} \times \mathbb{R}^{nm} \times \mathbb{R}^{pq} \to \mathbb{R}$ associated to problem (16) by $h(x(t), p(t), u(t)) = p(t)^\top f(x(t), u(t))$ and denote the local minimizer for problem (16) by $(x^*(t), u^*(t))$. Then there exists an arc $p : [0,T] \to \mathbb{R}^{nm}$ and a scalar η equal to 0 or 1 satisfying the following conditions (Pontryagin's Maximum Principle (Clarke (2013))):

- the nontriviality condition:
$$(\eta, p(t)) \neq 0 \quad {}^\forall t \in [0,T], \tag{17}$$
- the transversality condition:
$$-p(T) \in \eta \nabla l_f(x^*(T)) + N_e^L(x^*(T)), \tag{18}$$
- the adjoint equation for almost every $t \in [0,T]$
$$-\dot{p}(t) = D_x h(x^*(t), p(t), u^*(t)), \tag{19}$$
- the maximum condition for almost every $t \in [0,T]$
$$h(x^*(t), p(t), u^*(t)) = \sup_{u\in\omega} h(x^*(t), p(t), u). \tag{20}$$

Since ψ_{nm}^{-1} exists, we obtain the Hamiltonian function associated to $h(x^*(t), p(t), u^*(t))$ as follows:
$$H(X^*(t), P(t), U^*(t)) = \mathrm{Tr}\left(P^\top(t) F(X^*(t), U^*(t))\right), \tag{21}$$
which satisfies (11), (12), (13), and (14), where $X^* = \psi_{nm}^{-1}(x^*)$, $U^* = \psi_{pq}^{-1}(u^*)$, $P = \psi_{nm}^{-1}(p)$. This completes the proof. □

Lemma 2. *Define a set*
$$E \triangleq \{A \in \mathbb{R}^{n\times n} : A_{i,j} \leq \alpha_{i,j}, \quad (i,j) \in \mathcal{I}\}, \tag{22}$$
and fix any $\gamma \in E$, where $\mathcal{I} \subset \mathbb{N}^{n\times n}$ is a set of positions of elements of A for which inequality constraints are given. Then any $\delta \in N_E^L(\gamma)$ satisfies
$$\delta_{i,j}(\gamma_{i,j} - \alpha_{i,j}) = 0 \quad {}^\forall (i,j) \in \mathcal{I}, \tag{23}$$
$$\delta_{i,j} \geq 0 \quad {}^\forall (i,j) \in \mathcal{I}, \tag{24}$$
$$\delta_{i,j} = 0 \quad {}^\forall (i,j) \notin \mathcal{I}. \tag{25}$$

Proof. Fix any $\hat{A} \in E$ and $\hat{a} = \psi_{nn}(\hat{A})$, where ψ_{nn} is from (15). Then we obtain a set e satisfying $\hat{a} \in e$ as follows:
$$e \triangleq \{a \in \mathbb{R}^{n^2} : a_j \leq \alpha'_j, \quad j \in \mathcal{I}'\}, \tag{26}$$
where $\alpha' = \psi_{nn}(\alpha)$ and $\mathcal{I}' \subset \mathbb{N}^{n^2}$ is a set corresponding to $\mathcal{I}$. Take any $\gamma' \in e$, then we have
$$\delta'_i(\gamma'_i - \alpha'_i) = 0 \quad {}^\forall i \in \mathcal{I}', \tag{27}$$
$$\delta'_i \geq 0 \quad {}^\forall i \in \mathcal{I}', \tag{28}$$
$$\delta'_i = 0 \quad {}^\forall i \notin \mathcal{I}' \tag{29}$$
for all $\delta' \in N_E^L(\gamma')$ (Ikeda and Kashima (2022)). Finally, we obtain (23), (24), (25) where $\delta = \psi_{nn}^{-1}(\delta')$ and $\gamma = \psi_{nn}^{-1}(\gamma')$. □

Theorem 1. *Let $G_c^*(t)$ and $V^*(t)$ be a local optimal solution of Problem 2. Assume that*
$$q_j(t) \triangleq b_j^\top e^{A^\top(T-t)} \frac{\partial K(G_c^*(T))}{\partial G_c^*(T)} e^{A(T-t)} b_j$$
and $q_i(t) - q_j(t)$ is not constant on $[0,T]$ for all $i, j \in \{1,2,\ldots,p\}$. Then any solution to Problem 2 takes only the values in the binary set {0,1} almost everywhere.

Proof. We first reformulate Problem 2 into a form to which Lemma 1 is applicable. The value $\|v_j\|_{L^1}$ is equal to the final state $y_j(T)$ of the system $\dot{y_j}(t) = v_j(t)$ with $y_j(0) = 0$. Define $Y(t) \triangleq \mathrm{diag}(y(t))$ and matrices $X(t), \bar{V}(t), \bar{A}, \bar{B}$ by
$$X(t) \triangleq \begin{bmatrix} G_c(t) & O_{n\times p} \\ O_{p\times n} & Y(t) \end{bmatrix}, \ \bar{V}(t) \triangleq \begin{bmatrix} V(t) & O_{p\times p} \\ O_{p\times p} & V(t) \end{bmatrix},$$
$$\bar{A} \triangleq \begin{bmatrix} A & O_{n\times p} \\ O_{p\times n} & O_{p\times p} \end{bmatrix}, \ \bar{B} \triangleq \begin{bmatrix} B & O_{n\times p} \\ O_{p\times p} & I_p \end{bmatrix}.$$
Then, Problem 2 is equivalently expressed as follows:
$$\begin{aligned} \min_v \quad & J(V) = -L_f(X(T)) \\ \text{s.t.} \quad & \dot{X}(t) = \bar{A}X(t) + X(t)\bar{A}^\top + \bar{B}\bar{V}(t)\bar{B}^\top, \\ & X(0) = O_{(n+p)\times(n+p)}, \quad X(T) \in E, \\ & v(t) \in \Omega \quad {}^\forall t \in [0,T], \end{aligned} \tag{30}$$
where $L_f(X(T)) = K(G_c(T))$, $E = \{X(T) : y_j(T) \leq \alpha_j \quad {}^\forall j \in \{1,2,\ldots,p\}\}, \Omega = \{v(t) : v(t) \in [0,1]^p, \|v(t)\|_{l^1} \leq \beta\}$. This is an optimal control problem to which Lemma 1 is applicable. We define the Hamiltonian function H associated to problem (30) by
$$H(X, P, V) = \mathrm{Tr}\left(P^\top(\bar{A}X(t) + X(t)\bar{A}^\top + \bar{B}\bar{V}(t)\bar{B}^\top)\right).$$
We define two matrices as follows:
$$X^*(t) \triangleq \begin{bmatrix} G_c^*(t) & O_{n\times p} \\ O_{p\times n} & Y^*(t) \end{bmatrix}, \quad \bar{V}^*(t) \triangleq \begin{bmatrix} V^*(t) & O_{p\times p} \\ O_{p\times p} & V^*(t) \end{bmatrix}. \tag{31}$$
Then $(X^*(t), \bar{V}^*(t))$ is the local minimizer of problem (30) because of the equivalence between Problem 2 and problem (30), and there exists a scalar η equal to 0 or 1 and a matrix $P : [0,T] \to \mathbb{R}^{n\times n}$ satisfying the conditions (11), (12), (13), (14). It follows from (13) that
$$-\dot{P}(t) = \bar{A}^\top P(t) + P(t)\bar{A},$$
which leads to
$$\begin{aligned} P(t) &= e^{\bar{A}^\top(T-t)} P(T) e^{\bar{A}(T-t)} \\ &= \begin{bmatrix} e^{A^\top(T-t)} P^{(11)}(T) e^{A(T-t)} & e^{A^\top(T-t)} P^{(12)}(T) \\ P^{(21)}(T) e^{A(T-t)} & P^{(22)}(T) \end{bmatrix}, \end{aligned} \tag{32}$$
where
$$P(t) = \begin{bmatrix} P^{(11)}(t) & P^{(12)}(t) \\ P^{(21)}(t) & P^{(22)}(t) \end{bmatrix} \tag{33}$$

with $P^{(11)}(t) \in \mathbb{R}^{n\times n}$ and $P^{(22)}(t) \in \mathbb{R}^{p\times p}$. Note that

$$\begin{bmatrix} -P^{(11)}(T) + \eta \dfrac{\partial K(G_c^*(T))}{\partial G_c^*(T)} & -P^{(12)}(T) \\ -P^{(21)}(T) & -P^{(22)}(T) \end{bmatrix} \in N_E^L(X^*(T))$$

by (12), then we have

$$P^{(22)}_{j,j}(T)(y_j(T) - \alpha_j) = 0 \quad j = \{1, 2, \dots, p\}, \tag{34}$$

$$P^{(22)}_{j,j}(T) \le 0 \quad j = \{1, 2, \dots, p\}, \tag{35}$$

$$P^{(22)}_{i,j}(T) = 0 \quad {}^\forall(i, j) \in \{(i, j) : i \neq j\}, \tag{36}$$

$$-P^{(11)}(T) + \eta \frac{\partial K(G_c^*(T))}{\partial G_c^*(T)} = O_{n\times n}, \tag{37}$$

$$P^{(12)}(T) = O_{n\times p}, \quad P^{(21)}(T) = O_{p\times n}, \tag{38}$$

from Lemma 2. Substituting these into (32), we get

$$P(t) = \begin{bmatrix} e^{A^\top(T-t)} \eta \dfrac{\partial K(G_c^*(T))}{\partial G_c^*(T)} e^{A(T-t)} & O_{n\times p} \\ O_{p\times n} & P^{(22)}(T) \end{bmatrix}. \tag{39}$$

Then, we have

$$\begin{aligned} \mathrm{Tr}\left(P^\top(t)\bar{B}\bar{V}(t)\bar{B}^\top\right) &= \mathrm{Tr}\left(\bar{B}^\top P^\top(t)\bar{B}\bar{V}(t)\right) \\ &= \mathrm{Tr}\left(\begin{bmatrix} B^\top P^{(11)\top}(t) B & O_{p\times p} \\ O_{p\times p} & P^{(22)}(t) \end{bmatrix} \bar{V}(t)\right) \\ &= \sum_{j=1}^{p} \left(\eta q_j(t) + P^{(22)}_{j,j}(T)\right) v_j(t). \end{aligned}$$

It follows from (14) that

$$v^*(t) = \arg\max_{v\in\Omega} \sum_{j=1}^{p} \left(\eta q_j(t) + P^{(22)}_{j,j}(T)\right) v_j. \tag{40}$$

We here claim that $\eta = 1$. Indeed, if $\eta = 0$, $P^{(22)}(T) \neq O_{p\times p}$ follows from (11), i.e., there exists some j that satisfies

$$P^{(22)}_{j,j}(T) < 0, \quad y_j^*(T) = \alpha_j. \tag{41}$$

Hence, from (40) and (41), we have $v_j^*(t) = 0$ for all $t \in [0, T]$, i.e., $y_j^*(T) = \|v_j^*\|_{L^1} = 0$. This contradicts to (41). Thus, $\eta = 1$. From the assumption, it is easy to verify that

1) we have $q_j(t) + P^{(22)}_{j,j}(T) \neq 0$ almost everywhere for all $j = \{1, 2, \dots, p\}$,
2) there exists $j_k : [0, T] \to \{1, 2, \dots, p\}$, $k = 1, 2, \dots, p$, such that
$q_{j_1(t)}(t) + P^{(22)}_{j_1(t),j_1(t)}(T) > \cdots > q_{j_p(t)}(t) + P^{(22)}_{j_p(t),j_p(t)}(T)$
almost everywhere.

Hence, we find

$$v_j^*(t) = \begin{cases} 1 & \text{if} \quad j \in \Xi_1(t) \cap \Xi_2(t), \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

for almost every $t \in [0, T]$, where

$$\Xi_1(t) \triangleq \{j_1(t), j_2(t), \dots, j_\beta(t)\},$$

$$\Xi_2(t) \triangleq \{k \in \{1, 2, \dots, p\} : q_{j_k(t)}(t) + P^{(22)}_{j_k(t),j_k(t)}(T) > 0\}.$$

This completes the proof. □

The following theorem is the main result, which shows the equivalence between Problem 1 and Problem 2.

Theorem 2 (equivalence)**.** *Assume that $q_j(t)$ and $q_i(t) - q_j(t)$ is not constant on $[0, T]$ for all $i, j \in \{1, 2, \dots, p\}$. Denote the set of all solutions of Problem 1 and Problem 2 by $\mathcal{V}_0^*$ and $\mathcal{V}_1^*$, respectively. If the set $\mathcal{V}_1^*$ is not empty, then we have $\mathcal{V}_0^* = \mathcal{V}_1^*$.*

Proof. Denote any solution of Problem 2 by $\hat{v} \in \mathcal{V}_1^*$. It follows from Theorem 1 that $\hat{v}(t) \in \{0, 1\}^p$ almost everywhere. Note that the null set $\cup_{j=1}^p \{t \in [0, T] : \hat{v}_j(t) \notin \{0, 1\}\}$ does not affect the cost, and hence we can adjust the variables so that $\hat{v}(t) \in \{0, 1\}^p$ on $[0, T]$, without loss of the optimality. We have

$$\|\hat{v}(t)\|_{l^1} = \|\hat{v}(t)\|_{l^0}, \quad \|\hat{v}_j\|_{L^1} = \|\hat{v}_j\|_{L^0}$$

for all j. Since $\hat{v} \in \mathcal{V}_1$, we have $\|\hat{v}(t)\|_{l^0} \le \beta$ and $\|\hat{v}_j\|_{L^0} \le \alpha_j$ for all t and j. Thus, $\hat{v} \in \mathcal{V}_0$. Then,

$$J(\hat{v}) \le \max_{v\in\mathcal{V}_0} J(v) \le \max_{v\in\mathcal{V}_1} J(v) = J(\hat{v}), \tag{43}$$

where the first relation follows from $\hat{v} \in \mathcal{V}_0$, the second relation follows from $\mathcal{V}_0 \subset \mathcal{V}_1$, and the last relation follows from $\hat{v} \in \mathcal{V}_1^*$. Hence, we have

$$J(\hat{v}) = \max_{v\in\mathcal{V}_0} J(v), \tag{44}$$

which implies $\hat{v} \in \mathcal{V}_0^*$. Hence, $\mathcal{V}_1^* \subset \mathcal{V}_0^*$ and $\mathcal{V}_0^*$ is not empty.

Next, take any $\tilde{v} \in \mathcal{V}_0^*$. Note that $\tilde{v} \in \mathcal{V}_1$, since $\mathcal{V}_0^* \subset \mathcal{V}_0 \subset \mathcal{V}_1$. In addition, it follows from (44) that $J(\tilde{v}) = J(\hat{v})$. Therefore, $\tilde{v} \in \mathcal{V}_1^*$, which implies $\mathcal{V}_0^* \subset \mathcal{V}_1^*$. This gives $\mathcal{V}_0^* = \mathcal{V}_1^*$. □

4. CONCLUSION

In this paper, we discussed the equivalence between the sparsity constrained controllability metrics maximization problems and their convex relaxation. The proof is based on the matrix-valued Pontryagin maximum principle applied to the controllability Lyapunov differential equation. The existence of optimal solutions and computational cost are currently under investigation.

REFERENCES

Athans, M. (1967). The matrix minimum principle. *Information and Control*, 11, 592–606.

Clarke, F. (2013). *Functional Analysis, Calculus of Variations and Optimal Control*, volume 264. Springer Science & Business Media.

Ikeda, T. and Kashima, K. (2018). Sparsity-constrained controllability maximization with application to time-varying control node selection. *IEEE Control Systems Letters*, 2(3), 321–326.

Ikeda, T. and Kashima, K. (2022). Sparse control node scheduling in networked systems based on approximate controllability metrics. *IEEE Transactions on Control of Network Systems*.

Ikeda, T., Sakurama, K., and Kashima, K. (2021). Multiple sparsity constrained control node scheduling with application to rebalancing of mobility networks. *IEEE Transactions on Automatic Control*.

Olshevsky, A. (2014). Minimal controllability problems. *IEEE Transactions on Control of Network Systems*, 1(3), 249–258.

Pasqualetti, F., Zampieri, S., and Bullo, F. (2014). Controllability metrics, limitations and algorithms for complex networks. *IEEE Transactions on Control of Network Systems*, 1(1), 40–52.

Verriest, E. and Kailath, T. (1983). On generalized balanced realizations. *IEEE Transactions on Automatic Control*, 28(8), 833–844.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Controllability and observability of poset-causal systems

S. ter Horst * **J. Zeelie** **

* *Department of Mathematics, North-West University, Potchefstroom, 2520 South Africa (Sanne.TerHorst@nwu.ac.za).*
** *Department of Mathematics, North-West University, Potchefstroom, 2520 (e-mail: 24698245@nwu.ac.za).*

Abstract: Concepts of controllability and observability have been defined for a class of decentralized systems known as coordinated linear systems. The classical duality result does not extend to these systems. In the present paper, we generalize these notions of controllability and observability to poset-causal systems. We introduce the dual system associated with a poset-causal system and extend the classical duality result using this notion of a dual system.

Keywords: Decentralized systems, poset-causal systems, controllability and observability, duality.

1. INTRODUCTION

In this paper we consider poset-causal systems introduced by Shah and Parrilo in Shah et al. (2008) and further developed in the papers Shah et al. (2009, 2011, 2013) and the PhD thesis Shah (2011).

Poset causal systems are decentralized systems that consist of interconnected subsystems, labeled $1, 2, \ldots, p$, which are modeled by a partial order $\succeq$ on the set $P = \{1, 2, \ldots, p\}$. Hence, the pair $\mathcal{P} = (P, \succeq)$ is a partially ordered set (poset). In this setting subsystem j can 'influence' subsystem i in case $i \succeq j$. We assume that each subsystem is locally given by an input-state-output model with an input space $\mathcal{U}_i = \mathbb{R}^{m_i}$, a state space $\mathcal{X}_i = \mathbb{R}^{n_i}$ and an output space $\mathcal{Y}_i = \mathbb{R}^{r_i}$, with $m_i, n_i, r_i \in \mathbb{Z}_+$. The output y_i and state x_i are determined by the states and inputs of subsystems that can 'influence' subsystem i via interconnected state space system equations

$$\dot{x}_i(t) = \sum_{j \in \uparrow i} A_{ij} x_j + \sum_{k \in \uparrow i} B_{ik} u_k, \quad x_i(0) = x_{i,0},$$
$$y_i(t) = \sum_{j \in \uparrow i} C_{ij} x_j + \sum_{k \in \uparrow i} D_{ik} u_k, \quad t \geq 0,$$

where $\uparrow i = \{j \mid j \succeq i\}$. Here $x_{i,0} \in \mathbb{R}^{n_i}$ is the initial state of subsystem i and $A_{ij} \in \mathbb{R}^{n_i \times n_j}$, $B_{ij} \in \mathbb{R}^{n_i \times m_j}$, $C_{ij} \in \mathbb{R}^{r_i \times n_j}$ and $D_{ij} \in \mathbb{R}^{r_i \times m_j}$ are given matrices whenever $j \succeq i$. In case $j \not\succeq i$, set A_{ij}, B_{ij}, C_{ij} and D_{ij} equal to zero matrices of appropriate sizes and define

$$A = [A_{ij}],\ B = [B_{ij}],\ C = [C_{ij}],\ D = [D_{ij}], \qquad (1)$$

where $i, j = 1, 2, \ldots, p$. Then the combined input, state and output signals

$$u(t) = (u_1(t), \ldots, u_p(t))^\intercal \in \mathcal{U} = \oplus_{i=1}^p \mathcal{U}_i,$$
$$x(t) = (x_1(t), \ldots, x_p(t))^\intercal \in \mathcal{X} = \oplus_{i=1}^p \mathcal{X}_i,$$
$$y(t) = (y_1(t), \ldots, y_p(t))^\intercal \in \mathcal{Y} = \oplus_{i=1}^p \mathcal{Y}_i,$$

with $\oplus$ indicating orthogonal sums, satisfy

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0 \in \mathcal{X},$$
$$y(t) = Cx(t) + Du(t), \quad t \geq 0,$$

for $x_0 = (x_{1,0}, \ldots, x_{p,0})^\intercal \in \mathcal{X}$. In the block partitioning (1), the entries A_{ij}, B_{ij}, C_{ij} and D_{ij} are zero matrices whenever $i \not\succeq j$. Since $\succeq$ is a partial order, this zero-pattern is invariant under block matrix multiplication, provided the block sizes are compatible for multiplication. Poset causal systems have system matrices whose block zero patterns are determined by a partial order. The proofs of the results as well as a more detailed expositions will be presented in ter Horst et al. (2021).

Concepts of controllability and observability have been defined for a subclass of poset-causal systems known as coordinated linear systems in Kemper et al. (2012). In the present paper we extend these concepts of controllability and observability to poset-causal systems. The partially ordered structures that underly coordinated linear systems have a stronger notion of transitivity, defined in Bart et al. (2018) as in-ultra transitive, which do not allow for a suitable notion of duality. In the last part of the current paper we introduce a notion of duality for poset-causal systems, and present duality results for the notions of controllability and observability defined here.

2. POSET CAUSAL SYSTEMS

In this section we give a more formal definition of poset-causal systems and introduce the dual of a poset-causal system. For this we require some preliminary definitions and results on order structures and matrices with associated block zero-patterns.

2.1 Order structures

A *partially ordered set*, or *poset*, is a pair $\mathcal{P} = (P, \succeq)$ with P a set and $\succeq$ a partial order on P. That is, $\succeq$ is a

⋆ This work is based on research supported in part by the National Research Foundation of South Africa (Grant Numbers 118513 and 127364).

binary relation on P which is reflexive, transitive and antisymmetric. For $i, j \in P$ we write $i \succ j$ if $i \succeq j$ and $i \neq j$. Also, $i \preceq j$ and $i \prec j$ mean $j \succeq i$ and $j \succ i$, respectively. In the sequel we will only consider finite posets, usually of the form $P = \{1, 2, \ldots, p\}$ for some positive integer p. Given a subset $R \subseteq P$ of a poset $\mathcal{P} = (P, \succeq)$ we define its downstream set $\downarrow R$ and its upstream set $\uparrow R$ as

$$\begin{aligned} \downarrow R &= \{i \in P \colon \exists\, j \in R \text{ such that } j \succeq i\} \quad \text{and} \\ \uparrow R &= \{i \in P \colon \exists\, j \in R \text{ such that } i \succeq j\}. \end{aligned} \tag{2}$$

In the case that R is a singleton, say $R = \{i\}$, we simply write $\downarrow i$ and $\uparrow i$.

Definition 1. Given a poset $\mathcal{P} = (P, \succeq)$, we define the *dual poset* as $\mathcal{P}_d = (P, \succeq_d)$ where

$$j \succeq_d k \iff k \succeq j$$

for each $j, k \in P$.

Given the dual poset $\mathcal{P}_d = (P, \succeq_d)$ of some poset $\mathcal{P} = (P, \succeq)$, we also define the upstream- and downstream sets in terms of the dual poset:

$$\uparrow_d i = \{j \in P : j \succeq_d i\} \text{ and } \downarrow_d i = \{j \in P : j \preceq_d i\}.$$

Clearly we have $\uparrow_d i = \downarrow i$ and $\downarrow_d i = \uparrow i$.

Example 2. Consider posets $\mathcal{P}_1$ and $\mathcal{P}_2$ with Hasse diagrams given by:

Then $\mathcal{G}^{\downarrow}_{\mathcal{P}_1}$ is the underlying digraph of a coordinated linear system with one coordinator and two followers. It represents a in-ultra transitive order. $\mathcal{G}^{\downarrow}_{\mathcal{P}_2}$ is the dual of $\mathcal{G}^{\downarrow}_{\mathcal{P}_1}$ and is not in-ultra transitive and hence is not the underlying poset of a coordinated linear system.

2.2 Block matrices with prescribed zero-patterns

In this paper we are interested in classes of block matrices with prescribed zero-patterns, which are not necessarily square, but are closed under (block) matrix multiplication, provided the sizes of the blocks are compatible. This requires us to introduce some notation.

Given some $n \in \mathbb{Z}_+$, we will say $\underline{n} = (n_1, n_2, \ldots, n_p) \in \mathbb{Z}^p_+$ is a partition of n if $|\underline{n}| := n_1 + n_2 + \ldots + n_p = n$. Let $\underline{n} = (n_1, n_2, \ldots, n_p) \in \mathbb{Z}^p_+$ and $\underline{m} = (m_1, m_2, \ldots, m_q) \in \mathbb{Z}^q_+$ be two given partitions. We will write $A = [G_{ij}] \in \mathbb{R}^{\underline{n} \times \underline{m}}$, in which case it is to be understood that $G_{ij} \in \mathbb{R}^{n_i \times m_j}$. In case $\underline{r} \in \mathbb{Z}^p_+$ is another partition, then matrices $G \in \mathbb{R}^{\underline{r} \times \underline{n}}$ and $H \in \mathbb{R}^{\underline{n} \times \underline{m}}$ are said to be compatible for block matrix multiplication GH. When only the lengths p of $\underline{n}$ and q of $\underline{m}$ are specified, we will sometimes speak of $p \times q$ block matrices.

By analogy of the incidence algebras studied in Davis (1970), we define block matrices with zero-pattern prescribed by a partial order in the following manner.

Definition 3. Given a poset $\mathcal{P} = (P, \succeq)$, with $P = \{1, \ldots, p\}$ and partitions $\underline{n}, \underline{m} \in \mathbb{Z}^p_+$, we define the *block incidence vector space* $\mathcal{I}^{\underline{n} \times \underline{m}}_{\mathcal{P}} \subseteq \mathbb{R}^{\underline{n} \times \underline{m}}$ as the subspace

$$\mathcal{I}^{\underline{n} \times \underline{m}}_{\mathcal{P}} = \{G = [G_{ij}] \in \mathbb{R}^{\underline{n} \times \underline{m}} \colon G_{ij} = 0 \text{ if } j \not\succeq i\}.$$

By arguments similar to those in Davis (1970) it follows that the block zero structure is preserved under block matrix multiplication when the matrices are compatible for block matrix multiplication.

Proposition 4. Let $\mathcal{P} = (P, \succeq)$ be a poset with p elements and let $\underline{n}, \underline{m}, \underline{r} \in \mathbb{Z}^p_+$. If $G \in \mathcal{I}^{\underline{r} \times \underline{n}}_{\mathcal{P}}$ and $H \in \mathcal{I}^{\underline{n} \times \underline{m}}_{\mathcal{P}}$, then the product GH is well-defined and $GH \in \mathcal{I}^{\underline{r} \times \underline{m}}_{\mathcal{P}}$. If $G \in \mathcal{I}^{\underline{n} \times \underline{n}}_{\mathcal{P}}$ and $\det(G) \neq 0$, then $G^{-1} \in \mathcal{I}^{\underline{n} \times \underline{n}}_{\mathcal{P}}$.

Throughout the paper we work with block compressions associated with subsets of P. Note that we have defined partitionings in such a way that zero entries are allowed. It will be convenient in this paper to define compressions by simply setting some of the entries in the partitionings equal to zero.

Definition 5. Let $P = \{1, \ldots, p\}$ and let $R, S \subset P$. Let $G \in \mathbb{R}^{\underline{n} \times \underline{m}}$ for partitions $\underline{n}, \underline{m} \in \mathbb{Z}^p_+$. Then $G(R, S)$ denotes the block matrix in $\mathbb{R}^{\underline{n}_R \times \underline{m}_S}$ where

$$\underline{n}_R = (n_{1,R}, \ldots, n_{p,R}) \in \mathbb{Z}^p_+, \text{ with } n_{j,R} = \begin{cases} 0 & \text{if } j \notin R \\ n_j & \text{if } j \in R \end{cases}$$

$$\underline{m}_S = (m_{1,S}, \ldots, m_{q,S}) \in \mathbb{Z}^q_+, \text{ with } m_{j,S} = \begin{cases} 0 & \text{if } j \notin S \\ m_j & \text{if } j \in S \end{cases}$$

and where $G(R, S) = [\widetilde{G}_{ij}]_{i,j=1,\ldots,p}$ is defined by

$$\widetilde{G}_{ij} = G_{ij} \text{ if } i \in R \text{ and } j \in S$$

and $\widetilde{G}_{ij}$ vacuous if $i \notin R$ or $j \notin S$.

If R is a singleton, say $R = \{i\}$, we write $A(i, S)$ and likewise we write $A(R, j)$ if $S = \{j\}$. For one-sided compressions, we follow Matlab notation, and write $A(:, S)$ in case $R = P$, or $A(R, :)$ in case $S = P$.

Theorem 6. Given a poset $\mathcal{P} = (P, \succeq)$ with p elements, partitionings $\underline{n}, \underline{m}, \underline{r} \in \mathbb{Z}^p_+$ and subsets $Q, S \subset P$, for any block matrices $G \in \mathbb{R}^{\underline{r} \times \underline{n}}$ and $H \in \mathcal{I}^{\underline{n} \times \underline{m}}_{\mathcal{P}}$ we have

$$(GH)(Q, S) = G(Q, R)H(R, S),$$

for any subset $R \subseteq P$ with $\downarrow S \subseteq R$, in particular $(GH)(Q, S) = G(Q, \downarrow S)H(\downarrow S, S)$.

Finally, we define the block identity matrix $I_{\underline{n}} \in \mathbb{R}^{\underline{n} \times \underline{n}}$ with respect to a partitioning $\underline{n} \in \mathbb{Z}^p_+$ as the block diagonal matrix in $\mathbb{R}^{\underline{n} \times \underline{n}}$ with identity matrices as diagonal blocks. Then $I_{\underline{n}}(:, S)$ can be viewed as the embedding of $\mathbb{R}^{\underline{n}_S}$ into $\mathbb{R}^{\underline{n}}$ and $I_{\underline{n}}(S, :)$ as the projection from $\mathbb{R}^{\underline{n}}$ onto $\mathbb{R}^{\underline{n}_S}$.

2.3 Poset causal systems

Definition 7. Let $\mathcal{P} = (P, \succeq)$ be a poset with $P = \{1, \ldots, p\}$. A poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$ (with underlying poset $\mathcal{P}$) is a linear time invariant system with structured system matrices

$$A \in \mathcal{I}^{\underline{n} \times \underline{n}}_{\mathcal{P}}, \quad B \in \mathcal{I}^{\underline{n} \times \underline{m}}_{\mathcal{P}}, \quad C \in \mathcal{I}^{\underline{r} \times \underline{n}}_{\mathcal{P}}, \quad D \in \mathcal{I}^{\underline{r} \times \underline{m}}_{\mathcal{P}}, \tag{3}$$

for $\underline{n}, \underline{m}, \underline{r} \in \mathbb{Z}^p_+$ and some initial state $x_0 \in \mathcal{X} = \mathbb{R}^{\underline{n}}$.

For a poset-causal systems $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$, with A, B, C, D as in (3), since $A(i, j) = 0$ if $j \not\succeq i$, for the transpose A^{T} we have $A^{\mathsf{T}}(j, i) = 0$ if $j \not\succeq i$, that is, if $i \not\preceq_d j$. Consequently, we have

$$A^{\mathsf{T}} \in \mathcal{I}^{\underline{n} \times \underline{n}}_{\mathcal{P}_d}, \quad B^{\mathsf{T}} \in \mathcal{I}^{\underline{m} \times \underline{n}}_{\mathcal{P}_d}, \quad C^{\mathsf{T}} \in \mathcal{I}^{\underline{n} \times \underline{r}}_{\mathcal{P}_d}, \quad D^{\mathsf{T}} \in \mathcal{I}^{\underline{m} \times \underline{r}}_{\mathcal{P}_d}.$$

This observation justifies the following definition of the dual system.

Definition 8. For a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$, we define its *dual poset-causal system* as

$$\Sigma_{\mathcal{P}_d} \sim (A^\intercal, C^\intercal, B^\intercal, D^\intercal).$$

Example 9. Consider the poset $\mathcal{P}_1$ in example 2. And consider a poset-causal system $\Sigma_{\mathcal{P}_1} \sim (A, B, C, 0)$ with $A \in \mathcal{I}_{\mathcal{P}_1}^{\underline{n}\times\underline{n}}$, $B \in \mathcal{I}_{\mathcal{P}_1}^{\underline{n}\times\underline{m}}$ and $C \in \mathcal{I}_{\mathcal{P}_1}^{\underline{r}\times\underline{n}}$ with $\underline{n} = (1,1,2)$, $\underline{m} = (1,1,1)$ and $\underline{r} = (1,1,1)$, given by

$$A = \left[\begin{array}{c|c|cc} 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right],\ B = \left[\begin{array}{c|c|c} 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 1 & 0 & 0 \\ \hline 0 & 0 & 1 \end{array}\right],\ C = \left[\begin{array}{c|c|cc} 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 \end{array}\right]$$

Then $\mathcal{X}_1 = \text{span}\{e_1\}$, $\mathcal{X}_2 = \text{span}\{e_2\}$ and $\mathcal{X}_3 = \text{span}\{e_3, e_4\}$. So that $\mathcal{X} = \text{span}\{e_1, e_2, e_3, e_4\} = \mathbb{R}^4$. The dual poset of $\mathcal{P}_1$ is $\mathcal{P}_d = \mathcal{P}_2$ given in example 2. And $\Sigma_{\mathcal{P}_d} \sim (A_d, B_d, C_d, 0)$ is given by:

$$A_d = A^\intercal = \left[\begin{array}{c|c|cc} 1 & 1 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right],\ B_d = C^\intercal = \left[\begin{array}{c|c|c} 1 & 1 & 1 \\ \hline 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right],$$

$$C_d = B^\intercal = \left[\begin{array}{c|c|cc} 1 & 0 & 1 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$$

Now $\Sigma_{\mathcal{P}_1}$ is a coordinated linear system and its dual system $\Sigma_{\mathcal{P}_d}$ is a poset-causal system which is not a coordinated linear system.

3. UPSTREAM CONTROLLABILITY AND DOWNSTREAM OBSERVABILITY

In this section we extend certain notions of controllability and observability introduced in Kemper et al. (2012) for coordinated linear systems to the setting of poset-causal systems. Consider a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$ with n-dimensional state space $\mathcal{X}$. The reachable subspace $\mathcal{R}$ and unobservable subspace $\mathcal{N}$ of $\mathcal{X}$ are defined as

$$\mathcal{R} = \text{Im}\,\mathcal{C}(A, B) \quad \text{and} \quad \mathcal{N} = \text{Ker}\,\mathcal{O}(C, A), \tag{4}$$

where

$$\mathcal{C}(A, B) = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix},\ \mathcal{O}(C, A) = \mathcal{C}(A^\intercal, C^\intercal)^\intercal.$$

For $i \in P$ we define the *i-downstream reachable set* $\mathcal{R}_i$ as

$$\mathcal{R}_i := \text{Im}\ \mathcal{C}(A(\downarrow i, \downarrow i), B(\downarrow i, i)) \subseteq \mathcal{X}(\downarrow i), \tag{5}$$

with $\mathcal{X}(\downarrow i)$ defined as

$$\mathcal{X}(\downarrow i) := \bigoplus_{j \in \downarrow i} \mathcal{X}_j.$$

Hence a vector $\xi \in \mathcal{X}(\downarrow i)$ is in $\mathcal{R}_i$ if it is reachable in the system

$$\dot{x}^{\downarrow i}(t) = A(\downarrow i, \downarrow i)x^{\downarrow i} + B(\downarrow i, i)u_i,$$

which includes states which can be reached by applying only the i^{th} input u_i. In such a case we say that ξ is *i-downstream reachable.*

The following theorem extends Lemma 3.2 in Kemper et al. (2012) to poset-causal systems.

Theorem 10. Given a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, 0, 0)$ with $A \in \mathcal{I}_{\mathcal{P}}^{\underline{n}\times\underline{n}}$ and $B \in \mathcal{I}_{\mathcal{P}}^{\underline{n}\times\underline{m}}$, the reachable space $\mathcal{R} \subseteq \mathcal{X}$ is given in terms of $i-$downstream reachable sets $\mathcal{R}_i$ as

$$\mathcal{R} = \sum_{i=1}^{p} I_{\underline{n}}(:, \downarrow i)\mathcal{R}_i.$$

Note that $\mathcal{R}_j \subseteq \mathcal{X}(\downarrow j)$ and that $\mathcal{X}_i \subset \mathcal{X}(\downarrow j)$ for each $i \in \downarrow j$. For each $j \in P$ and $i \in \downarrow j$ we define the following subspaces of $\mathcal{X}_i$:

$$\overline{\mathcal{R}}_i^j := \mathcal{X}_i \cap \mathcal{R}_j \quad \text{and} \quad \widetilde{\mathcal{R}}_i^j := P_{\mathcal{X}_i}\mathcal{R}_j.$$

Here $P_{\mathcal{X}_i}$ is the orthogonal projection onto $\mathcal{X}_i$. One can view $\overline{\mathcal{R}}_i^j$ as the set of local states $x_i \in \mathcal{X}_i$ that can be reached from a local input u_j in such a way that the other states downstream from j remain unaffected. The subspace $\widetilde{\mathcal{R}}_i^j$, on the other hand, is the set of local states $x_i \in \mathcal{X}_i$ that can be reached from a local input u_j while the other states downstream from subsystem j may also be affected. By definition of the subspaces, we directly get that:

$$\bigoplus_{i \in \downarrow j} \overline{\mathcal{R}}_i^j \subseteq \mathcal{R}_j \subseteq \bigoplus_{i \in \downarrow j} \widetilde{\mathcal{R}}_i^j. \tag{6}$$

Next we define subspaces $\overline{\mathcal{R}}$ and $\widetilde{\mathcal{R}}$ which respect the structure imposed by the poset $\mathcal{P}$:

$$\begin{aligned} \overline{\mathcal{R}} &:= \bigoplus_{j \in P} \overline{\mathcal{R}}_j \quad \text{and} \quad \widetilde{\mathcal{R}} := \bigoplus_{j \in P} \widetilde{\mathcal{R}}_j, \quad \text{where} \\ \overline{\mathcal{R}}_j &:= \sum_{i \in \uparrow j} \overline{\mathcal{R}}_j^i \quad \text{and} \quad \widetilde{\mathcal{R}}_j := \sum_{i \in \uparrow j} \widetilde{\mathcal{R}}_j^i. \end{aligned} \tag{7}$$

Definition 11. We call a poset-causal system $\Sigma_{\mathcal{P}}$ *upstream controllable* if $\overline{\mathcal{R}}_j = \mathcal{X}_j$ for each $j \in P$ and we say $\Sigma_{\mathcal{P}}$ is *weakly upstream controllable* if $\widetilde{R}_j = \mathcal{X}_j$ for each $j \in P$.

The following result explains the connection with classical controllability.

Theorem 12. For a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$ we have the following inclusions

$$\overline{\mathcal{R}} \subseteq \mathcal{R} \subseteq \widetilde{R}.$$

In particular, if $\Sigma_{\mathcal{P}}$ is upstream controllable, then it is controllable. If $\Sigma_{\mathcal{P}}$ is controllable, then it is weakly upstream controllable.

Example 13. Consider the poset-causal system $\Sigma_{\mathcal{P}_1} \sim (A, B, C, 0)$ given in example 9. We can calculate its reachable set $\mathcal{R}$ using (4) and its downstream reachable sets $\mathcal{R}_i$ for $i = 1, 2, 3$ using (5):

$$\begin{aligned} \mathcal{R} &= \text{span}\{e_1 + e_3, e_2, e_4\}, \quad \mathcal{R}_1 = \text{span}\{e_1 + e_3, e_2\} \\ \mathcal{R}_2 &= \text{span}\{e_2\} \quad \text{and} \quad \mathcal{R}_3 = \text{span}\{e_4\} \end{aligned}$$

We can now compute the spaces $\overline{\mathcal{R}}$ and $\widetilde{R}$ using (7):

$$\overline{\mathcal{R}} = \text{span}\{e_2, e_4\} \quad \text{and} \quad \widetilde{R} = \text{span}\{e_1, e_2, e_3, e_4\}.$$

From this we see that $\Sigma_{\mathcal{P}_1}$ is not upstream controllable nor is it controllable, but it is indeed weakly upstream controllable.

Next we define several parallel notions of observability for poset-causal systems.

For $i, j \in P$, define the subspaces

$$\mathcal{X}(\uparrow i) := \bigoplus_{j \in \uparrow i} \mathcal{X}_j \qquad \text{and} \qquad \mathcal{X}(P \backslash \uparrow i) := \bigoplus_{j \notin \uparrow i} \mathcal{X}_j.$$

Then clearly $\mathcal{X}(\uparrow i) \dotplus \mathcal{X}(P \backslash \uparrow i) = \mathcal{X}(P) = \mathcal{X}$. For $i \in P$, we define the i-upstream unobservable set $\mathcal{N}_i$ as

$$\mathcal{N}_i = \ker \mathcal{O}(C(i, \uparrow i), A(\uparrow i, \uparrow i)) \subseteq \mathcal{X}(\uparrow i).$$

Hence, a vector $\xi \in \mathcal{X}(\uparrow i)$ is in $\mathcal{N}_i$ if it is unobservable in the system

$$\begin{aligned} \dot{x}^{\uparrow i}(t) &= A(\uparrow i, \uparrow i)x^{\uparrow i}(t), \qquad x^{\uparrow i}(0) = \xi \\ y^{\uparrow i}(t) &= C(i, \uparrow i)x^{\uparrow i}(t). \end{aligned}$$

In such a case we say that ξ is i-upstream *unobservable*.

The following theorem generalizes Lemma 4.2 in Kemper et al. (2012) to poset-causal systems.

Theorem 14. Given a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, 0, C, 0)$ with $A \in \mathcal{I}_{\mathcal{P}}^{n\times n}$ and $C \in \mathcal{I}_{\mathcal{P}}^{r\times n}$, the unobservable space $\mathcal{N}$ is given in terms of i-upstream unobservable spaces as:

$$\mathcal{N} = \bigcap_{i=1}^{p} \left(\mathcal{N}_i \oplus \mathcal{X}(P\backslash\uparrow i)\right).$$

Recall that $\mathcal{N}_i \subseteq \mathcal{X}_{\uparrow i}$ and that $\mathcal{X}_j \subseteq \mathcal{X}_{\uparrow i}$ if $j \in \uparrow i$. For each $j \in \uparrow i$, we define

$$\overline{\mathcal{N}}_i^j := \mathcal{N}_i \cap \mathcal{X}_j \quad \text{and} \quad \widetilde{\mathcal{N}}_i^j := P_{\mathcal{X}_j}\mathcal{N}_i.$$

From these definitions, we immediately get the following inclusions:

$$\bigoplus_{j\in\uparrow i} \overline{\mathcal{N}}_i^j \subseteq \mathcal{N}_i \subseteq \bigoplus_{j\in\uparrow i} \widetilde{\mathcal{N}}_i^j, \tag{8}$$

In analogy with (7) we define the following structured subspaces of $\mathcal{X}$:

$$\begin{aligned} \overline{\mathcal{N}} &:= \bigoplus_{j\in P} \overline{\mathcal{N}}^j \quad \text{and} \quad \widetilde{\mathcal{N}} := \bigoplus_{j\in P} \widetilde{\mathcal{N}}^j, \quad \text{where} \\ \overline{\mathcal{N}}^j &:= \bigcap_{i\in\downarrow j} \overline{\mathcal{N}}_i^j \quad \text{and} \quad \widetilde{\mathcal{N}}^j := \bigcap_{i\in\downarrow j} \widetilde{\mathcal{N}}_i^j. \end{aligned} \tag{9}$$

Based on these spaces, we introduce the following notions of observability.

Definition 15. We call a poset-causal system $\Sigma_{\mathcal{P}}$ *downstream observable* if $\widetilde{\mathcal{N}} = \{0\}$ and *weakly downstream observable* if $\overline{\mathcal{N}} = \{0\}$.

In analogy with Theorem 12, we obtain the next theorem.

Theorem 16. For a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$ we have the following inclusions

$$\overline{\mathcal{N}} \subseteq \mathcal{N} \subseteq \widetilde{\mathcal{N}}.$$

In particular, if $\Sigma_{\mathcal{P}}$ is downstream observable, then it is observable. If $\Sigma_{\mathcal{P}}$ is observable, then it is weakly downstream observable.

4. DUALITY

For classical centralized systems, duality between controllability and observability is given by the following well known result (see for example proposition 2.21 in Dullerud et al. (2013)).

Theorem 17. The pair (A, B) is controllable if and only if the pair $(B^\intercal, A^\intercal)$ is observable.

We now investigate a connection between controllability of a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$ and observability of its dual system $\Sigma_{\mathcal{P}_d} \sim (A_d, B_d, C_d, D_d)$. We note that since $\uparrow_d i = \downarrow i$ for each $i \in P$, we have by definition that

$$C_d(i, \uparrow_d i) = B^\intercal(i, \downarrow i) = (B(\downarrow i, i))^\intercal.$$

From this, we can show that

$$\mathcal{X}_{\uparrow i} \ominus (\mathcal{R}_i)^d = \mathcal{N}_i \quad \text{and} \quad \mathcal{X}_{\downarrow i} \ominus (\mathcal{N}_i)^d = \mathcal{R}_i$$

for each $i \in P$, where $(\mathcal{R}_i)^d = \operatorname{Im} \mathcal{C}(A^d, B^d)$ and $(\mathcal{N}^i)^d = \ker \mathcal{O}(C^d, A^d)$. It is then possible to prove the next result.

Theorem 18. For a poset-causal system $\Sigma_{\mathcal{P}} \sim (A, B, C, D)$ and its dual $\Sigma_{\mathcal{P}_d} \sim (A^d, B^d, C^d, D^d)$, we have

$$(\overline{\mathcal{R}})^d = \widetilde{\mathcal{N}}^\perp, \quad (\widetilde{\mathcal{R}})^d = \overline{\mathcal{N}}^\perp, \quad (\overline{\mathcal{N}})^d = \widetilde{\mathcal{R}}^\perp, \quad (\widetilde{\mathcal{N}})^d = \overline{\mathcal{R}}^\perp.$$

In particular, $\Sigma_{\mathcal{P}}$ is upstream controllable (weakly upstream controllable) if and only if $\Sigma_{\mathcal{P}_d}$ is downstream observable (weakly downstream observable).

Example 19. For the dual system $\Sigma_{\mathcal{P}_d}$ given in example 9, we can now easily determine the subspaces $(\overline{\mathcal{N}})^d$, $\mathcal{N}^d$ and $(\widetilde{\mathcal{N}})^d$ using the duality result Theorem 18 and the subspace determined in example 13:

$$(\overline{\mathcal{N}})^d = \widetilde{\mathcal{R}}^\perp = \{0\},\ \mathcal{N}^d = \mathcal{R}^\perp = \operatorname{span}\{e_1 - e_3\} \quad \text{and}$$
$$(\widetilde{\mathcal{N}})^d = \overline{\mathcal{R}}^\perp = \operatorname{span}\{e_1, e_3\}$$

Hence in accordance with Theorem 18, we see that $\Sigma_{\mathcal{P}_d}$ is weakly downstream observable, but not observable or downstream observable.

ACKNOWLEDGEMENTS

This work is based on research supported in part by the National Research Foundation of South Africa (NRF) and the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS). Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF and CoE-MaSS do not accept any liability in this regard.

REFERENCES

H. Bart, T. Ehrhardt, and B. Silbermann, L-free directed bipartite graphs and echelon-type canonical forms, *Oper. Theory Adv. Appl.* **271** (2018), 75–117.

R.L. Davis, Algebras defined by patterns of zeros, *J. Combinatorial Theory* **9** (1970), 257–260.

G.E. Dullerud, F. Paganini, A Course in Robust Control Theory A Convex Approach, *Springer* (2013).

S. ter Horst and J. Zeelie, Realization theory for poset-causal systems: controllability, observability and duality, *Math. Control Signals Systems*, to appear.

P.L. Kempker, A.C.M. Ran, and J.H. van Schuppen, Controllability and observability of coordinated linear systems, *Linear Algebra Appl.* **437** (2012), 121–167.

L. Lessard, M. Kristalny, and A. Rantzer, On structured realizability and stabilizability of linear systems, 2013 American Control Conference, IEEE, 2013.

P. Shah, A partial order approach to decentralized control, PhD thesis, Massachusetts Institute of Technology, 2011.

P. Shah and P.A. Parrilo, H_2-optimal decentralized control over posets: a state-space solution for state-feedback, *IEEE Trans. Automat. Control* **58** (2013), 3084–3096.

P. Shah and P.A. Parrilo, An optimal controller architecture for poset-causal systems, *50th IEEE Conference on Decision and Control and European Control Conference*, IEEE, 2011.

P. Shah and P.A. Parrilo, A poset framework to model decentralized control problems, *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, IEEE, 2009.

P. Shah and P.A. Parrilo, A partial order approach to decentralized control, *47th IEEE Conference on Decision and Control*, IEEE, 2008.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The differences between port-Hamiltonian, passive and positive real descriptor systems

Karim Cherifi * **Hannes Gernandt** * **Dorothea Hinsen** *

* *Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin (e-mail: {cherifi, gernandt, hinsen}@math.tu-berlin.de).*

Abstract: For linear time-invariant descriptor systems it is well known that port-Hamiltonian systems are passive and that passive systems are positive real. In our contribution we study under which assumptions also the converse implications hold. We also study the relationship between passivity, KYP inequalities and a finite required supply.

Keywords: Linear systems, Descriptor systems, Passivity, Dissipativity, Port-Hamiltonian systems, KYP inequality, Positive real.

1. INTRODUCTION

Port-Hamiltonian (pH) systems have been increasingly used in recent years as a unified structured framework for energy based modeling of systems (van der Schaft and Jeltsema (2014); Jacob and Zwart (2012); Ortega et al. (2001); van der Schaft (2004); Beattie et al. (2019); Mehrmann and Unger (2022)). This type of formulation has gained increased interest from engineers and mathematicians due to its modeling flexibility and robustness properties.

Specifically, pH systems are of particular interest in coupled networks of systems and multiphysics simulation and control. This network coupling oftentimes imposes additional algebraic constraints on the system which naturally lead to *linear time-invariant descriptor systems* in state-space form presented as

$$\begin{aligned} \tfrac{d}{dt}Ex(t) &= Ax(t) + Bu(t), \quad x(0) = x_0, \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

where $u : \mathbb{R} \to \mathbb{K}^m$, $x : \mathbb{R} \to \mathbb{K}^n$, $y : \mathbb{R} \to \mathbb{K}^m$ are the *input*, *state*, and *output* of the system and $E, A \in \mathbb{K}^{n\times n}, B \in \mathbb{K}^{n\times m}$, $C \in \mathbb{K}^{m\times n}$, $D \in \mathbb{K}^{m\times m}$, and $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. The system (1) will be briefly denoted by $\Sigma = (E, A, B, C, D)$ and throughout it is assumed that the pair (E, A) is *regular* which means that $\lambda E - A$ is invertible for some $\lambda \in \mathbb{C}$.

It is well-known that pH descriptor systems are passive and that passive systems are positive real. Our aim is to provide sufficient conditions for the converse implications to hold.

2. NOTATIONS AND KNOWN RESULTS

For a pH descriptor systems the coefficients in (1) are assumed to have a special structure, see e.g. Beattie et al. (2018); van der Schaft (2013); van der Schaft and Maschke (2018); Morandin and Mehrmann (2019)

(pH) The system Σ is *port-Hamiltonian* if there exists $J, R, Q \in \mathbb{K}^{n\times n}$, $G, P \in \mathbb{K}^{n\times m}$, and $S, N \in \mathbb{K}^{m\times m}$ such that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} (J-R)Q & G-P \\ (G+P)^H Q & S+N \end{bmatrix}, \quad E^H Q \geq 0, \\ \begin{bmatrix} R & P \\ P^H & S \end{bmatrix} \geq 0, \quad J = -J^H, N = -N^H. \tag{2}$$

Here the quadratic function $\mathcal{H}(x) := \frac{1}{2}x^H E^H Q x$ is called the *Hamiltonian* and can be interpreted as the energy of the system. Furthermore, for a matrix $M \in \mathbb{K}^{n\times m}$, M^H denotes its conjugate transpose.

In this contribution, we will compare for descriptor systems the existence of a pH formulation with two other typically used system theoretic properties, namely, *passivity* (Pa) and *positive realness* (PR) which are introduced below.

It is well known that even non-linear and time-varying pH descriptor systems satisfy a certain power balance equation, see Morandin and Mehrmann (2019). This in particular implies the following *passivity* property of pH descriptor systems.

(Pa) The system Σ is *passive* if there exists $Q \in \mathbb{K}^{n\times n}$ such that $Q^H E = E^H Q \geq 0$ and $\mathcal{S}(x) = \frac{1}{2}x^H Q^H E x$ satisfies for all $T > 0$ and all consistent initial values x_0 and smooth functions x, u, y

$$\mathcal{S}(x(T)) - \mathcal{S}(x(0)) \leq \int_0^T y(\tau)^H u(\tau) d\tau. \tag{3}$$

The functions $\mathcal{S}$ satisfying the above conditions will be called *storage functions*.

The property (Pa) is hard to verify in practice since one would have to consider all possible solution trajectories. A better suited algebraic criterion for (Pa) is a linear matrix inequality which can be obtained by differentiation of (3). It was developed for ordinary systems, i.e. $E = I_n$ and independently by Kalman, Yakubovich and Popov and for descriptor systems e.g. in Zhang et al. (2002); Freund and Jarre (2004).

(KYP) The system Σ has a solution Q to the *generalized KYP inequality* if

$$\begin{bmatrix} -A^HQ - Q^HA & C^H - Q^HB \\ C - B^HQ & D + D^H \end{bmatrix} \geq 0, \quad E^HQ \geq 0. \quad (4)$$

In many applications only input-output data is given and hence an important question is whether we can decide if a system is pH from this data and even more, we want to obtain a pH representation (2) of the system. The typical approach is to use apply a Laplace transform to (1) which leads to the *transfer function*

$$\mathcal{T}(s) := C(sE - A)^{-1}B + D \quad (5)$$

that describes the input-output behavior in the frequency domain. It is well known for ordinary system that the passivity implies that its transfer function is *positive real*, see Anderson (1967); Anderson and Vongpanitlerd (1973).

(PR) The system Σ and the transfer function $\mathcal{T}$ are called *positive real* if $\mathcal{T}$ has no poles in $\mathbb{C}^+ := \{s \in \mathbb{C} | \text{Re } s > 0\}$ and $\mathcal{T}(s) + \mathcal{T}(s)^H \geq 0$ for all $s \in \mathbb{C}^+$;

Hence, if we want to obtain a pH formulation of a system, the system must be passive implying that its transfer function is positive real. If the data does not allow us to conclude (PR), e.g. due to measurement errors, one determines the nearest positive real transfer function Gillis and Sharma (2018). Therefore, the remaining task is to find a pH state space representation (2) of this positive real transfer function, i.e. one has to reconstruct the coefficients.

In the case of ODE systems which are controllable and observable, i.e. *minimal*, it is well known that (pH), (Pa), (PR) are equivalent. Hence our focus lies on the non-minimal case. For non-minimal ODE systems, a detailed study has been conducted in (Brogliato et al., 2007, Chapter 3) but without including pH systems. Another recent survey was given in Hughes and Branford (2021) see p. 59 therein for a discussion on unobservable and uncontrollable systems. The main goal is to study which implications between the aforementioned properties hold for descriptor systems, and to carve out for which implications the controllability and observability assumptions are crucial.

3. MAIN CONTRIBUTION

For descriptor systems the relations between (pH),(Pa), (KYP) and (PR) were already studied in numerous works Zhang et al. (2002); Freund and Jarre (2004); Masubuchi (2006); Camlibel and Frasca (2009); Reis and Stykel (2010); Reis et al. (2015); Reis and Voigt (2015). However to the best of our knowledge not all four properties have been investigated at the same time and oftentimes the minimality of the descriptor system is assumed.

As a first step, those of the aforementioned results which do not require minimality to obtain

$$\text{(pH)} \implies \text{(KYP)} \implies \text{(Pa)} \implies \text{(PR)}.$$

The following examples show that without further assumptions the converse implications are not true.

(i) The system $\dot{x} = -x + u, \quad y = 0$ has a solution to the KYP inequalities (4), but it is not pH.
(ii) The descriptor system $x = 0$, $y = x$ is passive but there is no solution to the KYP.
(iii) The system
$$\dot{x} = x, \quad y = x + u$$
is positive real but it is not passive.

Hence the remaining questions which we will answer in the talk are

(Q1) When do solutions to the KYP inequality lead to a pH formulation?
(Q2) When does passivity leads to solutions of the KYP inequality?
(Q3) Can every positive real transfer function be realized as a pH system?

The counter example (i) indicates that answer to question (Q1) is related to observability properties of the system. It was shown already for ODE systems in (van der Schaft, 2009, p. 55) that only those solutions Q to the KYP which satisfy

$$\ker Q \subseteq \ker A \cap \ker C \quad (6)$$

lead to a pH formulation. Conversely, the Q used in (pH) automatically satisfies (6).

We show that the same condition is true for descriptor systems. If the system is behaviorally observable then $\ker A \cap \ker C = \{0\}$ and hence the existence of a pH formulation is equivalent to the existence of invertible solutions to the KYP inequality.

The question (Q2) was already studied in Reis et al. (2015); Camlibel and Frasca (2009) where it was shown that passivity only guarantees the KYP inequality to hold on certain subspaces and as a consequence, we can only derive a pH formulation on a subspace. As a contribution, we derive for descriptor systems with index one a modified KYP inequality which can be solved for passive descriptor systems. If the system is in addition behaviorally observable then we derive a pH formulation of the system.

Regarding (Q3) we can explicitly construct a minimal realization as a pH system based on the well-known representation of positive real functions

$$\mathcal{T}(s) = M_1 s + \mathcal{T}_p(s)$$

for some $M_1 \geq 0$ and a proper positive real rational function $\mathcal{T}_p(s)$.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge valuable discussions with Volker Mehrmann, Arjan van der Schaft and Philipp Schulze.

REFERENCES

Anderson, B. (1967). A system theory criterion for positive real matrices. *J. Control*, 5(2), 171–182.
Anderson, B. and Vongpanitlerd, S. (1973). *Network Analysis and Synthesis*. Englewood Cliffs, NJ, USA: Prentice-Hall, Inc.
Beattie, C., Mehrmann, V., and van Dooren, P. (2019). Robust port-Hamiltonian representations of passive systems. *Automatica*, 100, 182–186.
Beattie, C., Mehrmann, V., Xu, H., and Zwart, H. (2018). Linear port-Hamiltonian descriptor systems. *Mathematics of Control, Signals, and Systems*, 30. doi: https://doi.org/10.1007/s00498-018-0223-3.

Brogliato, B., Lozano, R., Maschke, B., and Egeland, O. (2007). *Dissipative Systems Analysis and Control*. Springer Science+Business Media.

Camlibel, M. and Frasca, R. (2009). Extension of Kalman–Yakubovich–Popov lemma to descriptor systems. *Systems & Control Letters*, 58(12), 795–803. doi: https://doi.org/10.1016/j.sysconle.2009.08.010.

Freund, R. and Jarre, F. (2004). An extension of the positive real lemma to descriptor systems. *Optim. Methods Softw.*, 19(1), 69–87. doi:10.1080/10556780410001654232. URL `https://doi.org/10.1080/10556780410001654232`.

Gillis, N. and Sharma, P. (2018). Finding the nearest positive-real system. *SIAM J. Numer. Anal.*, 56(2), 1022–1047.

Hughes, T. and Branford, E. (2021). Dissipativity, reciprocity and passive network synthesis: from jan willems' seminal dissipative dynamical systems papers to the present day. *arXiv:2102.08855*.

Jacob, B. and Zwart, H. (2012). *Linear port-Hamiltonian systems on infinite-dimensional spaces*. Operator Theory: Advances and Applications, 223. Birkhäuser/Springer Basel AG, Basel CH.

Masubuchi, I. (2006). Dissipativity inequalities for continuous-time descriptor systems with applications to synthesis of control gains. *Systems & Control Letters*, 55(2), 158–164. doi: https://doi.org/10.1016/j.sysconle.2005.06.007.

Mehrmann, V. and Unger, B. (2022). Control of port-Hamiltonian differential-algebraic systems and applications. Technical report. ArXiv:2201.06590.

Morandin, R. and Mehrmann, V. (2019). Structure-preserving discretization for port-Hamiltonian descriptor systems. *IEEE 58th Conference on Decision and Control (CDC)*, 6863–6868.

Ortega, R., van der Schaft, A.J., Mareels, Y., and Maschke, B.M. (2001). Putting energy back in control. *Control Syst. Mag.*, 21, 18–33.

Reis, T., Rendel, O., and Voigt, M. (2015). The Kalman–Yakubovich–Popov inequality for differential-algebraic systems. *Linear Algebra Appl.*, 485, 153–193. doi:https://doi.org/10.1016/j.laa.2015.06.021.

Reis, T. and Stykel, T. (2010). Positive real and bounded real balancing for model reduction of descriptor systems. *Internat. J. Control*, 83(1), 74–88.

Reis, T. and Voigt, M. (2015). The Kalman–Yakubovich–Popov inequality for differential-algebraic systems: Existence of nonpositive solutions. *Systems & Control Letters*, 86, 1–8. doi:https://doi.org/10.1016/j.sysconle.2015.09.003.

van der Schaft, A.J. (2004). Port-Hamiltonian systems: network modeling and control of nonlinear physical systems. In *Advanced Dynamics and Control of Structures and Machines*, CISM Courses and Lectures, Vol. 444. Springer Verlag, New York, N.Y.

van der Schaft, A.J. (2009). Port-hamiltonian systems. In *Modeling and Control of Complex Physical Systems: The port-Hamiltonian approach*. Springer, Netherlands. doi:10.1007/978-3-642-03196-0.

van der Schaft, A.J. (2013). Port-Hamiltonian differential-algebraic systems. In A. Ilchmann and T. Reis (eds.), *Surveys in Differential-algebraic equations I*, Differential-algebraic Equations Forum, 173–226. Springer, Berlin.

van der Schaft, A.J. and Jeltsema, D. (2014). Port-Hamiltonian systems theory: An introductory overview. *Foundations and Trends in Systems and Control*, 1(2-3), 173–378.

van der Schaft, A.J. and Maschke, B. (2018). Generalized port-Hamiltonian DAE systems. *Systems Control Lett.*, 121, 31–37.

Zhang, L., Lam, J., and Xu, S. (2002). On positive realness of descriptor systems. *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications*, 49(3), 401–407.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Continuous Hedging Strategy for Power Market Using Financial Instruments on Electricity Price and Weather

Yuji Yamada* and Takuji Matsumoto**

* *Faculty of Business Sciences, University of Tsukuba, Tokyo 112-0012, Japan*
(e-mail: yuji@gssm.otsuka.tsukuba.ac.jp).

** *Faculty of Transdisciplinary Sciences, Kanazawa University, Ishikawa 920-1192, Japan*
(e-mail: mtakuji@staff.kanazawa-u.ac.jp)

Abstract: In this study, we develop a quantitative strategy for controlling cash-flow fluctuations of power utilities in electricity trading market using adequate financial instruments. In particular, we focus on hedging of thermal power generations and provide mixed positions of derivatives and forwards in a flexible manner, where we apply nonparametric regression techniques to find optimal payoff structure of derivatives and/or optimal units of forward contracts with fine granularity. An empirical backtest is conducted to illustrate our proposed hedging strategy.

Keywords: Forward contracts, derivatives, thermal power generators, hedging, nonparametric regressions.

1. INTRODUCTION

In the liberalized electricity market as illustrated in Figure 1, electricity retailing companies purchase spot electricity through the central power exchange and deliver it to their consumers, in which their volume for procurement needs to match future demand with uncertainty. On the other side of the market, power generation companies place sales orders and produce electricity based on the executed volume. In addition, due to the rapid increase of renewable power generation, the generators' supply volume depends on renewable energy sources (such as solar power) affected by weather conditions. In this situation, both electricity price and volume fluctuate in time, resulting in the risk of loss caused by high volatility of future cashflow. The objective of this study is to construct a quantitative strategy for reducing the cash-flow volatility in the electricity trading market based on financial instruments known as derivatives/forwards.

Derivatives/forwards are financial contracts whose payoffs depend on the values of underlying indexes at a future period and can be used as insurance purpose in electricity markets. For example, a power utility company (a power generator or a retailer) may be suffered from the loss caused by extremely high (or low) electricity price in the electricity market, but such risk can be avoided by purchasing an electricity derivative that compensates unexpected price difference. Also, weather or electricity derivatives can be used for the loss of volume changes resulting from unexpected weather condition (Bhattacharya et al. (2020), Oum et al. (2006), Matsumoto and Yamada (2021a), Coulon et al. (2013)).

In this study, we assume that such derivatives are offered by an insurance company and are constructed in flexible manner. In other words, we apply nonparametric regression techniques to find optimal payoff structure of derivatives and/or optimal units of forward contracts with fine granularity. Note that this work follows the results provided in our recent work of Yamada and Matsumoto (2021), but differs in the following aspects: Mixed positions of derivatives and forwards are constructed based on the balance between standard and specific products and the trading strategy is provided; Most recent data periods are covered to reflect the effect of COVID-19 pandemic and extreme price movement happened in January 2021 in Japan; Thermal power generations are particularly focused, where the thermal power generators' volume may approximately be given by demand minus renewable energy power generations, being largely influenced by solar radiations and other weather conditions.

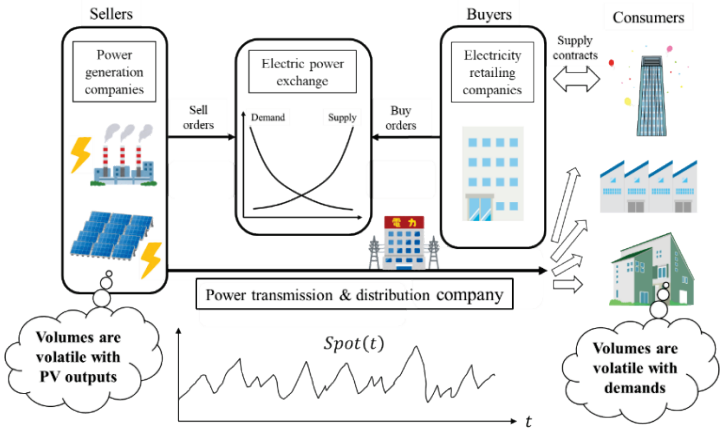


Figure 1. Electricity trading market.

2. PROBLEM SETTING

We introduce the notation and data used in our analysis as follows, where the data was observed in the Tokyo area, Japan from April 1st, 2016, (when the Japanese electricity market was fully liberalized) to December 31st, 2021:

$S_{t,m}$ [Yen/kWh]: JEPX spot price on day t delivering 1 kWh of electricity between hours m and $m+1$ (http://www.jepx.org/market/index.html).

$V_{t,m}$ [kWh]: Power generation of thermal generators between hours m and $m+1$ on day t. (https://www.tepco.co.jp/forecast/html/area_data-j.html)

$T_{t,m}$ [° C] : Temperature index in constructed using the electricity consumption-based weighted average of nine observation points in Tokyo area at hour m on day t. (https://www.data.jma.go.jp/gmd/risk/obsdl/)

$R_{t,m}$ [MJ/m^2]: Solar radiation index constructed using an installed capacity of local photovoltaics (PV) weighted average of seven observation points in Tokyo area between hours m and $m+1$ on day t. (https://www.data.jma.go.jp/gmd/risk/obsdl/)

Assume that there is a supply aggregator that compiles all the generation stacks from thermal generators and that the supply generator is willing to mitigate the fluctuations of total cash-flows for power generation at each period defined by $V_{t,m}S_{t,m}$. One measure of cash-flow variation is its variance, $\mathrm{Var}[V_{t,m}S_{t,m}]$, and in this study, we consider the problem of minimizing the cash-flow variance using forwards and derivatives. Such a strategy may be referred to as minimum variance hedging and has been investigated in energy markets (see e.g., Halkos and Tsirivis (2019) and references therein).

At the heart of hedging is to replicate the target cash-flow using another cash-flow defined by payoff of a portfolio of forward/derivative contracts. In this work, we construct a mixed position of forward and derivative contracts on weather and electricity indexes for thermal generators (or the supply aggregator). To the end, we apply the following generalized additive model (GAM; Hastie and Tibshirani (1990), Wood (2017)) for each $m \in \{0, \dots, 23\}$:

$$V_{t,m}S_{t,m} = \delta_m(t)S_{t,m} + \gamma_m(t)T_{t,m} + h_m(R_{t,m}) + Calendar_m(t) + \epsilon_{t,m} \quad (1)$$

where δ_m and γ_m are yearly cyclical spline functions that modelled with by-variables, $S_{t,m}$ and $T_{t,m}$ (see Wood (2017)), h_m is a smoothing spline function and $\epsilon_{t,m}$ is a residual satisfying zero mean condition, $\overline{\epsilon_{t,m}} = 0$. Note that radiation derivatives (term $h_m(R_{t,m})$) are applied daytime only for $m \in \{8, \dots, 15\}$ and assume that $h_m \equiv 0$ for $m \notin \{8, \dots, 15\}$.

In (1), $Calendar_m(t)$ contains day of week, long-term, and seasonal trends as

$$Calendar_m(t) = \beta_1 Mon_t + \cdots + \beta_6 Sat_t + \beta_7 Holidays_t + Seasonal(t) + Longterm(t) + c \quad (2)$$

where $Mon_t, \dots, Sat_t$, and $Holidays_t$ denote day-of-week and holiday dummy variables that take $Mon_t = 1$ if the day of t is Monday or $Mon_t = 0$ otherwise, and so on. $Seasonal(t)$ denotes a yearly cyclical smoothing spline function and reflects the seasonal trend in $V_{t,m}S_{t,m}$, whereas $Longterm(t)$ is a smoothing spline function (e.g., a cubic spline function) of the day variable t. In GAM (1), all the spline functions and coefficients are estimated to minimize the so-called penalized residual sum of squares (PRSS) which corresponding to minimizing the sample variance of $\epsilon_{t,m}$ with smoothing conditions (Matsumoto and Yamada (2021a), Yamada and Matsumoto (2021)), in which the estimated coefficients and spline functions differ by hour m, but we omit specifying this dependence for brevity.

Note that $V_{t,m}S_{t,m}$ defines the cash-inflow for power generators whereas $\delta_m(t)S_{t,m} + \gamma_m(t)T_{t,m} + h_m(R_{t,m}) + Calendar_m(t)$ in (1) the cash-outflow when constructing the hedge position. In this case, the hedge portfolio may be constructed by taking short positions of electricity and temperature forwards for $\delta_m(t)$ and $\gamma_m(t)$ units, radiation derivatives with payoff function $h_m(R_{t,m})$, zero coupon bonds with face value $Calendar_m(t)$. Then, solving GAM (1) yields minimum variance hedging with forwards and derivatives and $\epsilon_{t,m}$ gives a hedge error.

3. EMPIRICAL BACKTESTS

In this section, we apply our proposed minimum variance hedging for power generators (or equivalently, the supply aggregator), in which we estimate optimal spline functions and other required parameters in (1) based on the in-sample data to evaluate its out-of-sample performance using the out-of-sample data. The data periods are given as follows:

In-sample: April 1st, 2016, to December 31st, 2020.
Out-of-sample: January 1st to December 31st, 2021.

In this study, GAMs were estimated using R 4.0.5 (https://www.R-project.org/) and the package mgcv (Wood (2021)) to obtain the series of smoothing spline functions, wherein the smoothing parameter is calculated by the generalized cross-validation criterion. All figures are plotted using MATLAB 2021a (MathWorks, Inc., Natick, MA, USA).

We have estimated smooth functions and coefficients of GAM (1) using the data of in-sample-period, and executed out-of-sample simulations by substituting the data observed in the out-of-sample period into the regression equation to compute predicted values of GAM (1). Such predicted values do not necessarily provide forecast values obtained in the past period because they contain explanatory variables to be measured at the same time as the target value is observed. However, plotting the predicted values of GAM together with the realized values of $V_{t,m}S_{t,m}$ may help us intuitively grasp how the hedge is performed in the out-of-sample period. The blue and red lines in Figure 2 denote the realized and predicted values of $V_{t,m}S_{t,m}$ for $m = 13$ (1-2 pm) in the out-of-sample period, although they seem to be almost overlapped in these plots. Note that, around 1-2 pm, the electricity demand tended to take its maximum in a day, whereas the electricity price was not the highest due to the effect of solar power generation. In the current electricity market, the 30 minutes electricity price tends to take its highest value early evening. This is because the electricity demand still remains high, but the solar power generation almost disappears around that time. In fact, the spot electricity price recorded extremely high values around 4-9pm in the middle of January 2021 (which were about 25 times higher than monthly average price of January 2020). Such an extreme movement was observed in Figure 2 in that period.

As mentioned earlier in this section, although the predicted values given by regression equation in the right-hand side of (1) are not forecast values determined in the in-sample period,

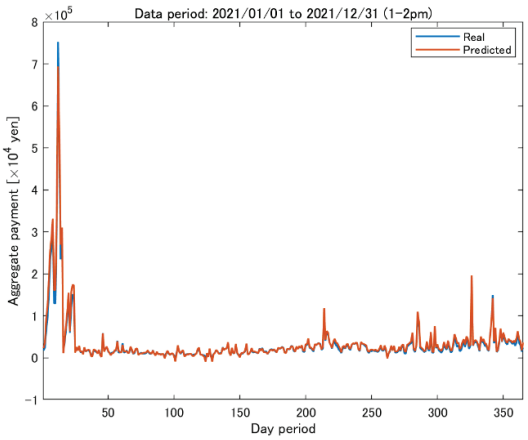


Figure 2. Realized vs. Predicted values of $V_{t,m}S_{t,m}$ in the out-of-sample period when $m = 14$ (1-2pm).

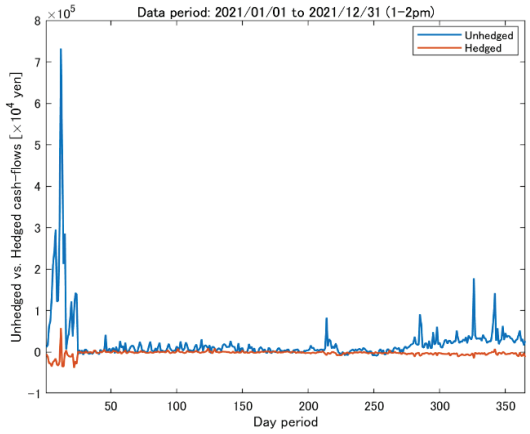


Figure 3. Unhedged vs. Hedged cash-flows in the out-of-sample period when (1-2pm).

the gap of two lines in Figure 2 provides hedge errors between $V_{t,m}S_{t,m}$ and the value of hedge portfolio consisting of forwards, derivatives, and bonds. To understand the effects of forwards and derivatives against the cash-flow fluctuation without these products, we consider a reference model using calendar trend only using GAM as

$$V_{t,m}S_{t,m} = Calendar_m(t) + \epsilon_{t,m}^{ref} \qquad (3)$$

where $\epsilon_{t,m}^{ref}$ is a residual satisfying zero mean condition, $\overline{\epsilon_{t,m}^{ref}} = 0$. In (3), $Calendar_m(t)$ may be thought of the face value of zero-coupon bond maturing on day t corresponding to the payment of debt from the income $V_{t,m}S_{t,m}$ for power generators. If the gap between $V_{t,m}S_{t,m}$ and $Calendar_m(t)$ is large, they may suffer from finding another source of financing. This is the idea to introduce financial instruments where the cash-flow fluctuations defined the residual terms are minimized using forwards and derivatives based on GAMs.

The blue and the red lines in Figure 3 compare unhedged vs. hedged cash-flows in the out-of-sample period for $m = 13$ (1-2pm). These lines were obtained by substituting the observed variables in (1) and (3) and plotting the residuals $\epsilon_{t,m}$ and $\epsilon_{t,m}^{ref}$ in the out-of-sample period for the hedged (red line) and the unhedged (blue line) cash-flows, respectively. Compared to the unhedged cash-flows, we see that the fluctuations of hedged cash-flows are significantly reduced by combining derivatives and forwards in these out-of-sample simulations.

3.3 Performance evaluation

We investigate the accuracy of the hedge in terms of the following Variance Reduction Rates (VRRs) and Normalized Mean Absolute Errors (NMAEs) for out-of-sample data:

$$VRR := \frac{\mathrm{Var}\left[\epsilon_{t,m}^{out}\right]}{\mathrm{Var}\left[\epsilon_{t,m}^{ref}\right]}, \qquad NMAE := \frac{\overline{\left|\epsilon_{t,m}^{out}\right|}}{\overline{\left|\epsilon_{t,m}^{ref}\right|}} \qquad (4)$$

where $\epsilon_{t,m}^{out}$ and $\epsilon_{t,m}^{ref}$ denote hedge errors corresponding to the residual of GAMs (1) and (3) with out-of-sample data. In (4), VRR and NMAE are defined by the ratios for improvement by applying derivatives and forwards compared to unhedged errors in the out-of-sample case. The bottom blue lines in Figures 4 and 5 provide VRRs and NMAEs for different values of $m = 0,1,\dots,23$ using hedged cash-flows with electricity and temperature forwards and radiation derivatives, i.e., the out-of-sample hedge errors for GAM (1). The red lines are the ones with electricity and temperature forwards and the yellow with electricity forwards only. Note that the radiation derivatives were applied for $m = 8,\dots,15$ (8am-4pm), the blue and the red lines take the same values outside this range. From these figures, we see that the introduction of temperature forwards reduced the cash-flow fluctuation early morning and daytime significantly, whereas the improvement in the evening was not observed almost at all. On the other hand, the introduction of radiation derivatives always improves the hedge performance.

4. CONCLUSION

In this paper, we have considered hedging of thermal power generators and provided mixed positions of forwards and derivatives on electricity price and weather indexes. The key findings of our analysis are summarized below.

1. Although unhedged cash-flows for power generators were unstable and had extreme price and volume fluctuations, they were shown to be stabilized by using financial instruments of our study. The fluctuations of hedged cash-flows were significantly improved by combining derivatives and forwards even in the out-of-sample simulations.

2. In contrast with power retailers' volume, the thermal generators' volume does not only depend on consumers' demand but also on the renewable power generation and others. The larger the amount of solar radiation, the lower the electricity price and the volume for thermal power generators (when demand is fixed). Our analysis showed that the radiation derivative is effective to reduce the cash-flow fluctuation resulting from uncertainty in solar power generation.

3. The cyclic trends in the units of electricity and the temperature forwards possessed in the hedge portfolio were modelled by combining cyclic splines with by-variables in GAMs and were shown to provide significant

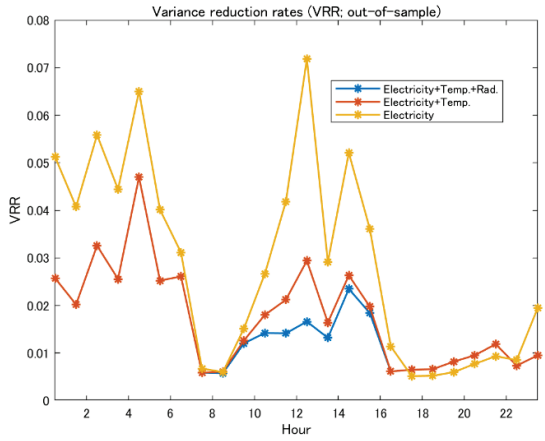


Figure 4. VRRs in the out-of-sample period.

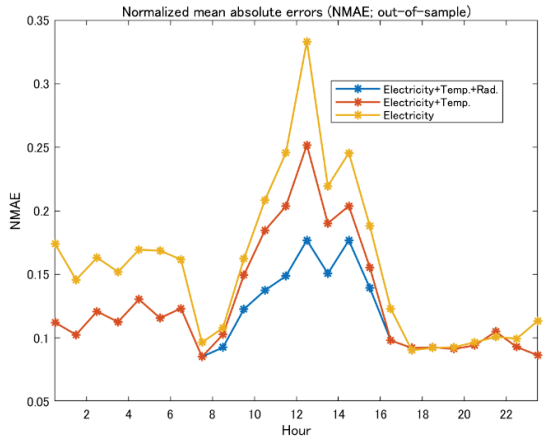


Figure 5. Normalized MAEs in the out-of-sample period.

hedge effects in terms of out-of-sample VRRs and NMAEs. Since electricity and temperature forwards are considered standardized products, showing their effectiveness is particularly important.

Compared to other electricity utility businesses, there is a variety of risks for thermal generators other than wholesale electricity price-volume fluctuations and the uncertainty of solar power generation. For example, volatility of fossil fuel energy prices is extremely high and the energy prices including oil and natural gas have increased currently (see e.g., https://www.eia.gov/, accessed on February 14th, 2022). Also, further investments in thermal power plants are becoming more difficult due to the regulations for controlling global warming gas emissions and many countries have been promoting a renewable energy policy. However, it is still fair to say that majority of electricity power generation depend on fossil fuel thermal power, and we need to consider risk management techniques for power generators from a cash-flow management perspective as well. In this sense, it is important to develop financial instruments and related strategy for reducing financial risk for power generators, which may lead to a smoother transition to renewable power as a core supply technology.

In practice, the interpretability of results is often important as well as accuracy, and we have taken a statistical approach based on GAM in this study, rather than other machine learning techniques. In this context, we have discussed the effectiveness of GAM for PV forecast in Matsumoto and Yamada (2021b), in comparison with four machine learning techniques and shown that the GAM-based PV forecast model flexibly express the tangled trend structure inherent in time series data and has an advantage not only in interpretability but also in improving forecast accuracy. A further investigation including a comparison of hedging models based on other machine learning techniques is interesting for future study.

ACKNOWLEDGEMENT

This work was funded by a Grant-in-Aid for Scientific Research (A) 20H00285, Grant-in-Aid for Challenging Research (Exploratory) 19K22024, and Grant-in-Aid for Young Scientists 21K14374 from the Japan Society for the Promotion of Science (JSPS).

REFERENCES

Bhattacharya, S., Gupta, A., Kar, K., and Owusu, A. (2020). Risk management of renewable power producers from co-dependencies in cash flows. *European Journal of operational research*, 283(3), 1081-1093.

Coulon, M., Powell, W. B. and Sircar, R. (2013). A model for hedging load and price risk in the Texas electricity market. *Energy Economics*, 40, 976-988.

Halkos, G. E. and Tsirivis, A. S. (2019). Energy commodities: A review of optimal hedging strategies. *Energies*, 12(20), 3979.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall: Boca Raton, FL, USA.

Matsumoto, T., and Yamada, Y. (2021). Comprehensive and Comparative Analysis of GAM-Based PV Power Forecasting Models Using Multidimensional Tensor Product Splines against Machine Learning Techniques. *Energies*, 14(21), 7146.

Matsumoto, T. and Yamada, Y. (2021). Simultaneous hedging strategy for price and volume risks in electricity businesses using energy and weather derivatives. *Energy Economics*, 95, 105101.

Oum, Y., Oren, S., and Deng, S. (2006). Hedging quantity risks with standard power options in a competitive wholesale electricity market. *Naval Research Logistics (NRL)*, 53(7), 697-712.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd edition, Chapman and Hall: New York, NY, USA.

Wood, S.N. (2021). Package 'mgcv' v. 1.8–38. https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

Yamada, Y., and Matsumoto, T. (2021). Going for Derivatives or Forwards? Minimizing Cashflow Fluctuations of Electricity Transactions on Power Markets. *Energies*, 14(21), 7311.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Targeted Harmonic Exploration with application to Robust Dual Control ⋆

Janani Venkatasubramanian * **Johannes Köhler** **
Julian Berberich * **Frank Allgöwer** *

* *Institute for Systems Theory and Automatic Control, University of Stuttgart, 70569 Stuttgart, Germany (email: {janani.venkatasubramanian, julian.berberich, frank.allgower}@ist.uni-stuttgart.de).*
** *Institute for Dynamic Systems and Control, ETH Zürich, Zürich CH-80092, Switzerland. (email: jkoehle@ethz.ch)*

Abstract: We present a novel targeted exploration strategy for the application of robust dual control. Unlike common greedy random exploration strategies considered in the related dual control literature, we introduce a targeted strategy in which the exploration inputs are a linear combination of sinusoids whose amplitudes are optimized based on an exploration criterion. Specifically, we leverage recent results on persistence of excitation using spectral lines to show how a (high probability) lower bound on the resultant persistence of excitation of the exploration data can be established. These results can be used to provide a priori lower bounds on the remaining model uncertainty after exploration. Given this exploration strategy and the corresponding uncertainty bounds, tools from robust control and gain-scheduling can be used to design a robust dual controller.

Keywords: Identification for control, Learning for control, Dual control

1. INTRODUCTION

Control inputs to an uncertain system should have a 'directing effect' to control the dynamical system and achieve a certain goal. Furthermore, the inputs should also have a 'probing' effect to learn the uncertainty in the system. In close association with these effects, simultaneous learning and control of uncertain dynamic systems has garnered research interest with the establishment of the *dual control* paradigm (Feldbaum, 1960). A detailed survey of dual control methods are provided by Filatov and Unbehauen (2000), and Mesbah (2018).

An approach to the dual control problem is to sequentially apply some probing input or exploration, and then a stabilizing controller informed by the gathered data. Recent methods by Umenberger et al. (2019), Barenthin and Hjalmarsson (2008), Ferizbegovic et al. (2019), and Iannelli et al. (2020) focus on *targeted* exploration and perform better than methods that use common greedy exploration. In particular, these methods consider that exploration should be targeted in the sense that the resulting uncertainty reduction in the model facilitates achieving a control goal and performance objective. The work of Umenberger et al. (2019) introduces a targeted exploration strategy that excites the system to reduce uncertainty specifically in a way that it minimizes the worst-case cost. This approach relies on a high probability uncertainty bound on system parameters which can be approximately predicted and thus optimized depending on the exploration input. Developing on the works of Barenthin and Hjalmarsson (2008), and Umenberger et al. (2019), a dual control strategy is proposed by Ferizbegovic et al. (2019) that minimizes the worst-case cost achieved by a robust controller that is synthesized with reduced model uncertainty. The approach by Iannelli et al. (2020) extends the exploration strategy by Umenberger et al. (2019) to a more realistic finite horizon problem setting that captures the trade-offs between exploration and exploitation better.

The exploration methods proposed by Venkatasubramanian et al. (2020), Barenthin and Hjalmarsson (2008), Umenberger et al. (2019), Ferizbegovic et al. (2019), and Iannelli et al. (2020), rely on state-feedback and an additional optimized Gaussian noise term for exploration. To tractably compute the predicted uncertainty bound associated with the parameter estimates after exploration, the empirical covariance is approximated by the worst-case state covariance. These approximations fail to provide a priori guarantees of excitation on the exploration inputs. Hence, we propose harmonic exploration inputs in the form of a linear combination of sinusoids of specific frequencies and reduce uncertainty in a targeted fashion with the goal of guaranteed control performance. This choice is also supported in literature, where it was established that the robust optimal control input can be expressed with appropriately chosen amplitudes and frequencies of the sinusoids (Rojas et al., 2007). Additionally, by suitably

⋆ F. Allgöwer is thankful that his work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016 and under grant 468094890. F. Allgöwer acknowledges the support by the Stuttgart Center for Simulation Science (SimTech). Janani Venkatasubramanian thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting her.

influencing the spectral content of inputs, these inputs provide the advantage of guaranteed persistence of excitation (Sarker et al., 2020). The spectral information of the exploration inputs can be used to determine a lower uncertainty bound on the persistence of excitation of the exploration data, which can in turn be used to design a robust controller that provides some desired performance guarantees. Regarding the application of targeted exploration in dual control, our exploration strategy can be employed in our recently proposed gain-scheduling based dual control approach (Venkatasubramanian et al., 2020). The gain scheduling based approach gives rise to a linear matrix inequality (LMI) based design with closed-loop performance guarantees.

2. PROBLEM FORMULATION

Notation The transpose of a matrix $A \in \mathbb{R}^{n\times m}$ is denoted by $A^\top$. The positive definiteness of a matrix $A \in \mathbb{R}^{n\times n}$ is denoted by $A = A^\top \succ 0$. The conjugate transpose of a matrix $A \in \mathbb{C}^{n\times m}$ with complex entries is denoted by A^{H}. The operator $\mathrm{vec}(A)$ stacks the columns of A to form a vector. The operator $\mathrm{diag}(A_1, ..., A_n)$ creates a block diagonal matrix by aligning the input matrices $A_1, ..., A_n$ along the diagonal starting with A_1 in the upper left corner. The critical value of the Chi-squared distribution with n degrees of freedom and probability p is denoted by $\chi^2_n(p)$. The space of square-summable sequences is denoted by ℓ_2. A unit sphere of dimension d is denoted by $\mathcal{S}^{d-1}$. The expected value of a random variable X is denoted by $\mathbb{E}[X]$. Given a sequence $\{x_k\}_{k=0}^{T-1}$, the discrete Fourier transform (DFT) of the sequence is denoted by $\mathbf{x}(e^{j\omega}) = \sum_{k=0}^{T-1} x_k e^{-j2\pi k\omega}$ where $\omega \in \Omega_T$ and $\Omega_T := \{0, 1/T, ..., (T-1)/T\}$. For a vector $x \in \mathbb{R}^n$, the Euclidean norm is denoted by $\|x\| = \sqrt{x^\top x}$. For a matrix $M \in \mathbb{C}^{m\times n}$, the largest singular value is denoted by $\|M\|$. A random variable $X \in \mathbb{R}^d$ that is normally distributed with mean μ and variance Σ is denoted by $X \sim \mathcal{N}(\mu, \Sigma)$. A random variable $X \in \mathbb{R}$ is said to be sub-Gaussian with variance proxy σ^2, i.e., $X \sim \mathsf{subG}(\sigma^2)$, if $\mathbb{E}[X] = 0$ and its moment generating function satisfies $\mathbb{E}[e^{sX}] \leq e^{\left(\frac{\sigma^2 s^2}{2}\right)}, \quad \forall s \in \mathbb{R}$. A random vector $X \in \mathbb{R}^d$ is said to be sub-Gaussian with variance proxy σ^2, i.e., $X \sim \mathsf{subG}(\sigma^2)$, if $\mathbb{E}[X] = 0$ and $u^\top X$ is sub-Gaussian with variance proxy σ^2 for any unit vector $u \in \mathcal{S}^{d-1}$.

Setting: Consider a discrete-time linear time-invariant dynamical system of the form

$$x_{k+1} = A_{\mathrm{tr}} x_k + B_{\mathrm{tr}} u_k + w_k, \quad w_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2 I) \qquad (1)$$

where $x_k \in \mathbb{R}^{n_x}$ is the state, $u_k \in \mathbb{R}^{n_u}$ is the control input, and $w_k \in \mathbb{R}^{n_x}$ is the normally distributed process noise. It is assumed that the realizations of w_k are independent and identially distributed (i.i.d.) and the state x_k is directly measurable. The true values of the system dynamics, A_{tr} and B_{tr}, are unknown.

Control goal and outline: The overarching goal of the proposed dual control strategy is to design a stabilizing state-feedback controller $u_k = Kx_k$ which ensures that the closed-loop system is stable while also satisfying some desired quadratic performance specifications, e.g., ℓ_2-gain with high probability (Scherer, 2001; Veenman and Scherer, 2014). Additionally, an initial estimate of the system parameters is available with potentially large uncertainty. Correspondingly, we propose a sequential dual control approach wherein a targeted exploration strategy is implemented first and is followed by the implementation of the parametrized state-feedback controller. The primary challenge is to encapsulate the dual effect of performance improvement through the process of exploration, and to tailor the exploration in a manner that is pertinent to performance improvement. Therefore, we simultaneously design a targeted exploration strategy and a parametrized state-feedback controller for the system in (1) in dependence of the future parameters/model estimate. The new parameter estimate, which influences the state-feedback control law K, is interpreted as a scheduling variable using tools from gain-scheduling (Venkatasubramanian et al., 2020). The preliminaries regarding uncertainty bounds for parameter estimation based on time-series data and spectral information are provided in Section 3. The exploration strategy and the corresponding uncertainty bound on the data obtained during exploration are elaborated in Section 4. The dual control strategy is summarized in Section 5.

3. UNCERTAINTY BOUND

This section discusses preliminary results that provide a data-dependent uncertainty bound on the parameter estimates, and a lower uncertainty bound on the resultant persistence of excitation based on spectral information, by Umenberger et al. (2019) and Sarker et al. (2020), respectively.

3.1 Data-dependent uncertainty bound

Given observed data $\mathcal{D} = \{x_k, u_k\}_{k=0}^{N-1}$ of length N, denote $\phi_k = \begin{bmatrix} x_k^\top & u_k^\top \end{bmatrix}^\top \in \mathbb{R}^{n_\phi}$ where $n_\phi = n_x + n_u$. The least squares estimate of the unknown parameters A_{tr} and B_{tr} is computed as:

$$(\hat{A}, \hat{B}) = \underset{A,B}{\arg\min} \sum_{k=0}^{N-1} ||x_{k+1} - Ax_k - Bu_k||_2^2. \qquad (2)$$

The following lemma provides a high probability credibility region for the uncertain system matrices.

Lemma 1. (Umenberger et al., 2019, Lemma 3.1) Given data set $\mathcal{D}$ and $0 < \delta < 1$, let $D = \frac{1}{\sigma_w^2 c_\delta} \sum_{k=1}^{N-1} \phi_k \phi_k^\top$ with $c_\delta = \chi^2_{n_x^2 + n_x n_u}(\delta)$. Suppose we have a uniform prior over the parameters (A, B), i.e., given $\theta = \mathrm{vec}([A, B])$, $p(\theta) \propto 1$. Then, $[A_{\mathrm{tr}}, B_{\mathrm{tr}}] \in \Theta$ with probability $1 - \delta$, where

$$\Theta := \left\{ A, B : \begin{bmatrix} (\hat{A} - A)^\top \\ (\hat{B} - B)^\top \end{bmatrix}^\top D \begin{bmatrix} (\hat{A} - A)^\top \\ (\hat{B} - B)^\top \end{bmatrix} \preceq I \right\}. \qquad (3)$$

The result of Lemma 1 is a data-dependent uncertainty bound that can be utilized to synthesize robust controllers similar to approaches by Umenberger et al. (2019), Ferizbegovic et al. (2019), Iannelli et al. (2020), and Venkatasubramanian et al. (2020). In Section 4, we use this to derive a suitable targeted exploration strategy. Since the dynamics are unknown in this setup, we compute an initial error bound of the form given in Lemma 1 from the following assumption.

Assumption 1. The parameters (A, B) have a uniform prior. Furthermore, an initial data set $\mathcal{D}_0 = \{\phi_k\}_{k=-T_0}^{-1}$ is available such that

$$D_0 \coloneqq \frac{1}{\sigma_w^2 c_\delta} \sum_{k=-T_0}^{-1} \phi_k \phi_k^\top \succ 0. \tag{4}$$

Initial estimates of the system parameters $\hat{A}_0$ and $\hat{B}_0$ can be derived from the data $\mathcal{D}_0$. The matrix D_0 quantifies the robust bound associated with these initial estimates for a given probability $1-\delta$ (cf. Lemma 1), and can be determined from $\mathcal{D}_0$ as given in (4). This initial data can be acquired through some random persistently exciting input. Through the targeted exploration process for T time steps (cf. Section 4), data $\mathcal{D}_T = \{\phi_k\}_{k=0}^{T-1}$ will be observed. The new estimates $\hat{A}_T$ and $\hat{B}_T$ will be computed from data $\mathcal{D}_0 \cup \mathcal{D}_T$ and made available at time T. The matrix $D_T \coloneqq D_0 + \frac{1}{\sigma_w^2 c_\delta} \sum_{k=0}^{T-1} \phi_k \phi_k^\top$ will quantify the uncertainty associated with the estimates $\hat{A}_T$ and $\hat{B}_T$ (cf. Lemma 1).

3.2 Finite excitation based on the theory of spectral lines

This subsection discusses parameter estimation results based on the theory of spectral lines (Bai and Sastry, 1985; Sarker et al., 2020), which deals with the analysis of frequency domain information that can be derived from time-series data. The finite-data uncertainty bounds in Lemma 1 are based on the matrix D, which corresponds to the following notion of finite excitation.

Definition 1. (Finite Excitation (Sarker et al., 2020)) A sequence $\{\phi_k\}_{k\geq 0,\dots T-1}$ is said to be finitely exciting from $k=0$ to $T-1$ if there exist constants $0 < \rho_1 \leq \rho_2$ such that

$$\rho_2 I \succeq \sum_{k=0}^{T-1} \phi_k \phi_k^\top \succeq \rho_1 I. \tag{5}$$

With finite data in the stochastic setting, the notion of a sub-Gaussian spectral line is introduced.

Definition 2. (Sub-Gaussian Spectral Line (Sarker et al., 2020)) A stochastic sequence $\{\phi_k\}_{k\geq 0,\dots T-1}$ is said to have a sub-Gaussian spectral line from $k=0$ to $T-1$ at a frequency $\omega_0 \in \Omega_T$ of amplitude $\bar{\phi}(\omega_0) \in \mathbb{C}^{n_\phi}$ and a radius $R > 0$ if

$$\frac{1}{T} \sum_{k=0}^{T-1} \phi_k e^{-j2\pi\omega_0 k} - \bar{\phi}(\omega_0) \sim \mathsf{subG}(R^2/T). \tag{6}$$

In order to establish the relationship between the spectral content of the input and finite excitation, the expected information matrix is first defined.

Definition 3. (Expected Information Matrix (Sarker et al., 2020)) Given a sequence $\{\phi_k\}_{k\geq 0}$ with L sub-Gaussian spectral lines at frequencies $\{\omega_1, ..., \omega_L\}$ from $k=0$ to $T-1$ with amplitudes $\{\bar{\phi}(\omega_1), ..., \bar{\phi}(\omega_L)\}$, the information matrix $\bar{\Phi}$ is defined as

$$\bar{\Phi} = \begin{bmatrix} | & \dots & | \\ \bar{\phi}(\omega_1) & \dots & \bar{\phi}(\omega_L) \\ | & \dots & | \end{bmatrix}. \tag{7}$$

In deterministic system identification, fast estimation of unknown parameters is made possible if $\bar{\Phi}$ is full rank and numerically well conditioned. Note that we consider n_ϕ spectral lines for simplicity, which is the smallest number required for finite excitation and results in a square matrix $\bar{\Phi}$. Subsequently, since $\omega_i \in \Omega_T$, $i = 1, ..., n_\phi$, we set $T \geq n_\phi$ and select n_ϕ frequencies from Ω_T. Furthermore, whether an input signal is finitely exciting or not can be ascertained from its spectral content. The following lemma, inspired by (Sarker et al., 2020, Proposition 2), presents a clear relationship between the spectral content of the signal and finite excitation by utilizing the expected information matrix.

Lemma 2. If a sequence $\{\phi_k\}_{k\geq 0,\dots T-1}$ has n_ϕ spectral lines at frequencies $\omega_1, ..., \omega_{n_\phi}$ from $k=0$ to $T-1$ with amplitudes $\{\bar{\phi}(\omega_1), ..., \bar{\phi}(\omega_{n_\phi})\}$, then, for any $\epsilon \in (0,1)$, ϕ_k satisfies

$$\frac{1}{T} \sum_{k=0}^{T-1} \phi_k \phi_k^\top \geq \frac{1}{n_\phi} \left((1-\epsilon) \bar{\Phi}^\top \bar{\Phi} + \left(1 - \frac{1}{\epsilon}\right) \tilde{W}^\top \tilde{W} \right). \tag{8}$$

Here, $\tilde{W}$ is a random matrix for which each column is sub-Gaussian with properties as derived in Section 4.2. $\tilde{W}$ captures the effect of the disturbances w in a suitably defined manner. Inequality (8) can be used to determine a lower bound $\overline{D}_T$ on the *informativity* of the exploration data before the process of exploration. Such bounds are crucial for robust dual strategies (cf. Venkatasubramanian et al. (2020)). Note that this lower bound depends on the information matrix $\bar{\Phi}$, which contains the amplitudes of the harmonic signals (cf. (7)), as well as on the size of the noise.

4. TARGETED EXPLORATION

In this section, we propose a targeted exploration strategy and the corresponding uncertainty bound on the data obtained through the process of exploration. Unlike strategies based on greedy random exploration, we introduce a targeted exploration strategy in the form of a linear combination of sinusoids in specific frequencies that explicitly allow for the shaping of model uncertainty. Furthermore, a lower bound on finite excitation of the exploration data can be achieved using spectral information of the exploration inputs (cf. Lemma 2). This lower bound is pivotal in the design of a robust dual controller (cf. Section 5).

4.1 Exploration strategy

The exploration input takes the form

$$u_k = \sum_{i=1}^{n_\phi} a_i \cos(\omega_i k), \quad (a_i \in \mathbb{R}^{n_u}),\ k = 0, ..., T-1 \tag{9}$$

where T is the exploration time and a_i is the amplitude of the sinusoidal input with frequency ω_i. The amplitude of the sub-Gaussian spectral line at frequency ω_i is denoted as $\overline{u}(\omega_i)$. Since the input is sinusoidal and deterministic, the amplitude of the spectral line is $|\overline{u}(\omega_i)| = a_i/2$, and the radius of the spectral line is 0. Denote $U_\mathrm{e} = \mathrm{diag}(U_1, ..., U_{n_\phi})$ where $U_i = \overline{u}(\omega_i)$. The exploration input is computed such that it excites the system sufficiently with minimal control energy based on the initial parameter estimates, i.e., $U_\mathrm{e}^\top U_\mathrm{e} \preceq \gamma_\mathrm{e}$ where the bound γ_e is desired to be small.

4.2 Parameter estimation bounds

For the system evolving under the exploration input as given in (9), the uncertainty bound on the parameters can be computed from the expected information matrix $\bar{\Phi}$ of the input. As a requisite to compute the uncertainty bound, it is necessary to establish the relationship between the spectral content of the observed state and the input. The transfer function from u_k to ϕ_k can be written as,

$$\boldsymbol{\phi}(e^{j\omega}) = \begin{bmatrix} (e^{j\omega}I - A_{\mathrm{tr}})^{-1}B_{\mathrm{tr}} \\ I_{n_u} \end{bmatrix} \mathbf{u}(e^{j\omega}) + \begin{bmatrix} (e^{j\omega}I - A_{\mathrm{tr}})^{-1} \\ 0 \end{bmatrix} \mathbf{w}(e^{j\omega}). \tag{10}$$

Given u_k as in (9) and from (Sarker et al., 2020, Lemma 1), ϕ_k has a sub-Gaussian spectral line from 0 to $T-1$ at frequency $\omega_0 \in \Omega_T$ with amplitude

$$\bar{\phi}(\omega_0) = \begin{bmatrix} (e^{j\omega_0}I - A_{\mathrm{tr}})^{-1}B_{\mathrm{tr}} \\ I_{n_u} \end{bmatrix} \bar{u}(\omega_0). \tag{11}$$

Denote

$$V_i = \begin{bmatrix} (e^{j\omega_i}I - A_{\mathrm{tr}})^{-1}B_{\mathrm{tr}} \\ I_{n_u} \end{bmatrix},\ V_{\mathrm{tr}} = [V_1, \cdots, V_{n_\phi}],$$
$$Y_i = \begin{bmatrix} (e^{j\omega_i}I - A_{\mathrm{tr}})^{-1} \\ 0 \end{bmatrix},\ Y_{\mathrm{tr}} = [Y_1, \cdots, Y_{n_\phi}],$$
$$W = \mathrm{diag}(\bar{w}(\omega_1), ..., \bar{w}(\omega_{n_\phi})).$$

Then,

$$\bar{\Phi} = [V_1U_1, \cdots, V_{n_\phi}U_{n_\phi}] = V_{\mathrm{tr}}U_{\mathrm{e}}.$$

One can show that Lemma 2 holds with $\tilde{W} = Y_{\mathrm{tr}}W$ and $\epsilon > 0$ from (10):

$$\frac{1}{T}\sum_{k=0}^{T-1}\phi_k\phi_k^\top \succeq \frac{1}{n_\phi}\left((1-\epsilon)U_{\mathrm{e}}^\top V_{\mathrm{tr}}^{\mathsf{H}}V_{\mathrm{tr}}U_{\mathrm{e}} + \left(1-\frac{1}{\epsilon}\right)\tilde{W}^\top\tilde{W}\right) \succeq \overline{D}_T \tag{12}$$

from which we can derive a lower bound on the parameter uncertainty. Recall that we want to derive a lower bound on the information in the exploration data D_T which involves the term $\sum_{k=0}^{T-1}\phi_k\phi_k^\top$. Here, V_{tr} is unknown, however, we can compute an estimate $\hat{V}$ using an estimate of the system parameters $\hat{A}_0$, $\hat{B}_0$. By suitably bounding $|V_{\mathrm{tr}} - \hat{V}|$, a valid lower bound $\overline{D}_T$ satisfying (12) can be computed. Hence, we can compute a lower bound on finite excitation as defined in (5) only from terms that are known to us. By appropriately choosing the input U_e, we can guarantee a certain finite excitation in the future.

5. ROBUST DUAL CONTROL

The proposed harmonic targeted exploration strategy with the associated lower bound in (12) can be combined with the dual control paradigm suggested by Venkatasubramanian et al. (2020). The exploration strategy and a robust state-feedback controller can be simultaneously designed such that applying the feedback after exploration ensures the satisfaction of some performance specifications with high probability. Initial estimates of the system parameters, with potentially large uncertainty, are available. During the process of exploration, the estimates $\hat{A}$ and $\hat{B}$ change due to new data. We utilize this information by modeling the true system in (1) as a linear parameter varying (LPV) system, where the estimates $\hat{A}$, $\hat{B}$ are measured online. Given uncertainty bounds on the initial estimates and the estimates obtained after exploration, the dual control problem can now be interpreted as a gain-scheduling problem as described by Venkatasubramanian et al. (2020). The new parameter estimates obtained after exploration are selected as a scheduling variable which influences the control law K, thereby encapsulating the dual effect of performance improvement through the process of exploration. The resulting dual controller with targeted harmonic exploration strategy can provide (high-probability) a priori performance bounds by utilizing the excitation guarantees derived in (12). Further details including a theoretical analysis of the proposed dual control strategy can be found in (Venkatasubramanian et al., 2022).

REFERENCES

Bai, E.W. and Sastry, S.S. (1985). Persistency of excitation, sufficient richness and parameter convergence in discrete time adaptive control. *Systems & control letters*, 6(3), 153–163.

Barenthin, M. and Hjalmarsson, H. (2008). Identification and control: Joint input design and H_∞ state feedback with ellipsoidal parametric uncertainty via LMIs. *Automatica*, 44(2), 543–551.

Feldbaum, A.A. (1960). Dual control theory. *Automation and Remote Control*, 21(9), 874–1039.

Ferizbegovic, M., Umenberger, J., Hjalmarsson, H., and Schön, T.B. (2019). Learning robust LQ-controllers using application oriented exploration. *IEEE Control Systems Letters*, 4(1), 19–24.

Filatov, N.M. and Unbehauen, H. (2000). Survey of adaptive dual control methods. *IEE Proceedings-Control Theory and Applications*, 147(1), 118–128.

Iannelli, A., Khosravi, M., and Smith, R.S. (2020). Structured exploration in the finite horizon linear quadratic dual control problem. In *Proc. 21st IFAC World Congress*, 959–964.

Mesbah, A. (2018). Stochastic model predictive control with active uncertainty learning: a survey on dual control. *Annual Reviews in Control*, 45, 107–117.

Rojas, C.R., Welsh, J.S., Goodwin, G.C., and Feuer, A. (2007). Robust optimal experiment design for system identification. *Automatica*, 43(6), 993–1008.

Sarker, A., Gaudio, J.E., and Annaswamy, A.M. (2020). Parameter estimation bounds based on the theory of spectral lines. *arXiv preprint arXiv:2006.12687*.

Scherer, C.W. (2001). LPV control and full block multipliers. *Automatica*, 37(3), 361–375.

Umenberger, J., Ferizbegovic, M., Schön, T.B., and Hjalmarsson, H. (2019). Robust exploration in linear quadratic reinforcement learning. In *Advances in Neural Information Processing Systems*, 15310–15320.

Veenman, J. and Scherer, C.W. (2014). A synthesis framework for robust gain-scheduling controllers. *Automatica*, 50(11), 2799–2812.

Venkatasubramanian, J., Köhler, J., Berberich, J., and Allgöwer, F. (2020). Robust dual control based on gain scheduling. In *Proc. 59th IEEE Conference on Decision and Control (CDC)*, 2270–2277. IEEE.

Venkatasubramanian, J., Köhler, J., Berberich, J., and Allgöwer, F. (2022). Sequential learning and control: targeted exploration for robust performance. *In preparation*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On error analysis of a closed-loop subspace model identification method ⋆

H k * K Ik **

* k k
k
** k k
k k

This article is a resubmission of the full paper that was accepted for the presentation at the MTNS 2020. This article reports error analysis and asymptotic variance of a closed-loop subspace model identification method for a system described with the output-error state-space representation. For details, since the procedure of the identification method includes the QR factorization of stacked data Hankel matrices, this study investigates asymptotic properties of block entries of the triangular matrix obtained from the QR factorization. The set of the block entries is separated into two components, namely, the signal-based component and the noise-based component. The contributions are to derive asymptotic properties of both components and to obtain the asymptotic covariance matrix of the vectorization of the noise-based component.

System identification, subspace methods, closed-loop identification, asymptotic properties, covariance matrices

1. INTRODUCTION

The authors have proposed a closed-loop subspace model identification method for a system described with the output-error state-space representation (Oku et al., 2006a,b). Since it utilizes the QR-factorization and its procedure resembles the procedure of the "MOESP" family very much, for convenience it is called CL-MOESP for the rest of this paper. A superiority of CL-MOESP is that it can directly estimate a system to be identified in closed-loop even when it is contaminated with the arbitrary colored output-error noise. Practical usefulness has been demonstrated by closed-loop system identification experiments on a cart-inverted-pendulum system (Oku and Ushida, 2009), a coaxial miniature helicopter (Matsuba et al., 2012; Kojio et al., 2014) and a quadrotor drone (Nakayama and Oku, 2018). Asymptotic properties and optimality of CL-MOESP have been discussed in part by Oku (2010). However, the analysis of consistency of the estimate and error, as well as the analysis of the asymptotic variance, is an important open problem.

With respect to the error analysis of CL-MOESP, this paper studies the estimation error of the triangular matrix obtained by the QR factorization of a stack of data Hankel matrices defined later. In details, focusing on significant block entries of the triangular matrix that are used to estimate the extended observability matrix and the coefficient matrices, the set of the block matrices is described using data Hankel matrices, and then it is separated into two components, namely, the signal-based component and the noise-based component. The contributions are as follows: 1) to prove that the signal-based component converges in probability to the matrix composed of meaningful structured matrices, each of which consists of the Markov parameters of the closed-loop system to be handled, 2) to prove that the noise-based component converges in probability to zero, and finally, 3) to derive the asymptotic covariance matrix of the vectorization of the noise-based component.

2. BRIEF REVIEW OF CL-MOESP

In this section, CL-MOESP in a general framework is briefly reviewed (Oku et al., 2006a,b).

Let us consider a closed-loop system depicted in Fig. 1. Suppose the plant P, which is to be identified, be a discrete-time linear time-invariant system described by the following minimal realization:

$$x_{k+1} = Ax_k + Bu_k, \qquad y_k = Cx_k + v_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$ denotes the state vector. Note that the order of the system, n, is unknown and it is to be estimated. Note also that the coefficients (A, B, C) are unknown and to be estimated up to a similarity transform. The output signal $y_k \in \mathbb{R}^l$ is measurable but it is contaminated with the colored noise $v_k \in \mathbb{R}^l$. The input to P, $u_k \in \mathbb{R}^m$, is generated by subtraction of the output filtered by a stabilizing controller, $K = K(q)$, from the external excitation signal r_k, that is,

$$u_k = r_k - K(q)y_k, \tag{2}$$

where q denotes the forward shift operator.

Let us assume the set point $d_k \equiv 0$. Then, $K(q)$ is not necessarily known [1].

⋆ This article is a resubmission of the full paper that was accepted at the MTNS 2020, with more concise notations adopted. This work was supported by JSPS KAKENHI Grant Number 18K04217.

[1] Also for the case where d is a constant reference signal, $K(q)$ is not necessarily known when the output is redefined as $y - d$.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

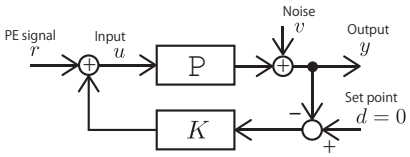


Fig. 1. Closed-loop system

Assume that, for $\forall k$ and $\forall i \geq 0$, x_k is independent of r_{k+i}. It means that the state at the current instant is influenced by the external excitation at relatively past instants. $\{r_k\}$ and $\{v_k\}$ are assumed to be uncorrelated each other in the sense that for $\forall i, \forall j \in \mathbb{Z}$, $\lim_{M\to\infty} \frac{1}{M} \sum_{k=1}^{M} v_{i+k} r_{j+k}^{\mathsf{T}} = 0$.

Given a sequence $\{u_k\}$, the block Hankel matrix, $\mathcal{U}_{i,s,j} \in \mathbb{R}^{ms\times j}$, is defined as

$$\mathcal{U}_{i,s,j} := \begin{bmatrix} u_i & u_{i+1} & \cdots & u_{i+j-1} \\ u_{i+1} & u_{i+2} & \cdots & u_{i+j} \\ \vdots & \vdots & & \vdots \\ u_{i+s-1} & u_{i+s} & \cdots & u_{i+s+j-2} \end{bmatrix}, \tag{3}$$

where the first subscript, i, is the same as the subscript on the top-left block element, and the others, s and j, mean that $\mathcal{U}_{i,s,j} \in \mathbb{R}^{ms\times j}$ is of s block rows and j columns. Note that $s > 0$ is a user-defined parameter which is chosen to be sufficiently larger than n. For sequences $\{y_k\}$, $\{r_k\}$ and $\{v_k\}$, the block Hankel matrices $\mathcal{Y}_{i,s,j}$, $\mathcal{R}_{i,s,j}$ and $\mathcal{V}_{i,s,j}$ are respectively defined in a manner similar to (3). These matrices are called the data Hankel matrices for the rest of this paper. For brevity's sake, the following notations with respect to the data Hankel matrices are introduced: for an integer N that is sufficiently larger than s,

$$\mathcal{R}_f := \mathcal{R}_{0,s,N}, \quad \mathcal{U}_f := \mathcal{U}_{0,s,N}, \quad \mathcal{Y}_f := \mathcal{Y}_{0,s,N},$$
$$\mathcal{R}_p := \mathcal{R}_{-s,s,N}, \quad \mathcal{U}_p := \mathcal{U}_{-s,s,N}.$$

Note that the subscriptions "f" and "p" respectively represents that the data Hankel matrices are made of relatively "future" and "past" data, respectively. Define $\mathcal{R} := \begin{bmatrix} \mathcal{R}_p^{\mathsf{T}} & \mathcal{R}_f^{\mathsf{T}} \end{bmatrix}$. For a sequence $\{x_k\}$, define

$$\mathcal{X}_{i,j} := \begin{bmatrix} x_i & x_{i+1} & \cdots & x_{i+j-1} \end{bmatrix}. \tag{4}$$

Given a quadruple of matrices (E, F, G, H) of appropriate sizes, define the notations as follows (Ikeda, 2014):

$$\mathcal{O}_i(G, E) := \begin{bmatrix} G^{\mathsf{T}} & (GE)^{\mathsf{T}} & \cdots & (GE^{i-1})^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}, \tag{5}$$

$$\mathcal{L}_i(E, F) := \begin{bmatrix} E^{i-1}F & \cdots & EF & F \end{bmatrix}, \tag{6}$$

$$\mathcal{T}_i(E, F, G, H) := \begin{bmatrix} H & & & 0 \\ GF & H & & \\ \vdots & \ddots & \ddots & \\ GE^{i-2}F & \cdots & GF & H \end{bmatrix}. \tag{7}$$

Using these notations, the extended observability matrix, $\mathcal{O}$, and the block-Toeplitz matrix made of the Markov parameters, $\mathcal{T}$, can be denoted as, respectively,

$$\mathcal{O} := \mathcal{O}_s(C, A), \qquad \mathcal{T} := \mathcal{T}_s(A, B, C, D).$$

Notice that from (1) $D = 0$ throughout this paper.

CL-MOESP is a solution to the following closed-loop subspace model identification problem.

Consider a closed-loop system depicted by Fig. 1. The problem is to estimate the order n of P to be identified and obtain the minimal realization (A, B, C) of P up to a similarity transform from the measurements of $\{r_k\}$, $\{u_k\}$ and $\{y_k\}$.

The procedure of CL-MOESP is as follows (Oku et al., 2006a,b):

(CL-MOESP). Suppose that a set of sampled data sequences $\{r_k\}$, $\{u_k\}$ and $\{y_k\}$ be obtained from a system identification experiment of the closed-loop system as depicted in Fig. 1. Then, a state-space model which represents the input/output relation of P can be obtained according to the procedures as follows:

1. Execute the QR factorization of the following matrix:
$$\begin{bmatrix} \mathcal{R} \\ \mathcal{U}_p \\ \mathcal{U}_f \\ \mathcal{Y}_f \end{bmatrix} = \begin{bmatrix} L_{11} & & & \\ L_{21} & L_{22} & & \\ L_{31} & L_{32} & L_{33} & \\ L_{41} & L_{42} & L_{43} & L_{44} \end{bmatrix} \begin{bmatrix} Q_1^{\mathsf{T}} \\ Q_2^{\mathsf{T}} \\ Q_3^{\mathsf{T}} \\ Q_4^{\mathsf{T}} \end{bmatrix} \tag{8}$$
2. Compute P and $\Upsilon^{\frac{1}{2}}$ as follows:
$$P := L_{21} - L_{21}L_{31}^{\mathsf{T}} \left(L_{31}L_{31}^{\mathsf{T}}\right)^{-1} L_{31} \tag{9}$$
$$\Upsilon^{\frac{1}{2}} := L_{41}P^{\mathsf{T}} \left(PP^{\mathsf{T}}\right)^{-\frac{1}{2}} \tag{10}$$
3. To estimate the extended observability matrix, $\mathcal{O}$, up to a similarity transform, execute singular value decomposition (SVD) of $\Upsilon^{\frac{1}{2}}$ and we have
$$\Upsilon^{\frac{1}{2}} = \begin{bmatrix} \hat{U} & \hat{U}^{\perp} \end{bmatrix} \begin{bmatrix} \hat{\Sigma} & \\ & \hat{\Sigma}^{\perp} \end{bmatrix} \begin{bmatrix} \hat{V}^{\mathsf{T}} \\ \left(\hat{V}^{\perp}\right)^{\mathsf{T}} \end{bmatrix},$$
where the diagonal matrix $\hat{\Sigma} \in \mathbb{R}^{n\times n}$ has n dominant singular values as its diagonal entries. Namely, the number of the dominant singular values can estimate the order of P. Note that the orthogonal matrix $\hat{U}$ is an estimate of $\mathcal{O}$ up to a similarity transform.
4. The set of estimates of the coefficients of a state-space representation of P, i.e., $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$, can be obtained according to the procedure similar in the paper of Verhaegen and Dewilde (1992). See appendix A of the literature (Oku and Ikeda, 2021) for the details.

3. PROBLEM SETTING AND NOTATIONS

The aim of this article is to study error analysis of the triangular matrix obtained from the QR factorization (8). Especially, since from (8), (9) and (10) the matrix with 3 significant block entries

$$\begin{bmatrix} L_{21}^{\mathsf{T}} & L_{31}^{\mathsf{T}} & L_{41}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}} \tag{11}$$

plays a key role in derivation of $\Upsilon^{\frac{1}{2}}$ and the subsequent procedures of CL-MOESP, we will investigate error analysis of (11) and derivation of its asymptotic covariance matrix.

Let us consider a closed-loop system depicted by Fig. 2. To simplify the problem, a constant gain feedback case will be considered. Let $u_k, r_k \in \mathbb{R}^m$ denote the input and the external excitation signal, respectively. $y_k, e_k \in \mathbb{R}^l$ denote the output and noise, respectively. The set point is assumed to be $d_k \equiv 0$ for $\forall k$. Suppose that P to be

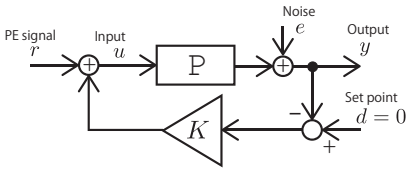


Fig. 2. A closed-loop system with constant gain feedback

identified is a n-th order LTI system of m inputs and l outputs with a minimal realization described by

$$x_{k+1} = Ax_k + Bu_k, \tag{12a}$$
$$y_k = Cx_k + e_k. \tag{12b}$$

where $x_k \in \mathbb{R}^n$ denotes the state vector of P. The input u_k is generated by subtraction of the output y_k multiplied by a constant feedback gain K from the external excitation signal r_k, that is,

$$u_k = r_k - Ky_k. \tag{13}$$

Suppose the closed-loop system be internally stable.

For the rest of the paper, adopt the following assumptions:

A1. $|\lambda_i(\bar{A})| < 1$, where $\bar{A} := A - BKC$.
A2. The noise $\{e_k\}$ is an unknown discrete-time Gaussian process with the mean $\mathrm{E}[e_k] = 0$ and the covariance $\mathrm{E}[e_k e_l^\mathsf{T}] = \Omega_{ee}\delta_{kl}$, which is not measurable.
A3. The external excitation signal $\{r_k\}$ is a known discrete-time Gaussian process with the mean $\mathrm{E}[r_k] = 0$ and the covariance $\mathrm{E}[r_k r_l^\mathsf{T}] = \Omega_{rr}\delta_{kl}$.
A4. $\{r_k\}$ and $\{e_k\}$ are uncorrelated with each other in the sense that, for $\forall i, \forall j \in \mathbb{Z}$, $\lim_{M\to\infty} \frac{1}{M}\sum_{k=1}^{M} e_{i+k} r_{j+k}^\mathsf{T} = 0$.
A5. For $\forall i \geq 0$, $\forall k$, x_k and r_{k+i} are independent of each other.
A6. The signals $\{x_k\}$, $\{r_k\}$, $\{u_k\}$, $\{y_k\}$ and $\{e_k\}$ are ergodic stationary processes.

The following notations are adopted: given i-successive sampled data of u from k, i.e., $\{u_k, \cdots, u_{k+i-1}\}$, define

$$u_i(k) := \begin{bmatrix} u_k^\mathsf{T} & \cdots & u_{k+i-1}^\mathsf{T} \end{bmatrix}^\mathsf{T}. \tag{14}$$

Note that $r_i(k)$, $y_i(k)$ and $e_i(k)$ are also defined in a similar manner to (14).

4. ERROR ANALYSIS OF CL-MOESP

x

Substitution of (13) into (12a) yields

$$x_{k+1} = \bar{A}x_k + Br_k + B_e e_k, \tag{15}$$

where $B_e := -BK$. Recursive use of (15) gives, for $i \geq 1$,

$$x_{k+i} = \bar{A}^i x_k + \mathcal{L}_i(\bar{A}, B) r_i(k) + \mathcal{L}_i(\bar{A}, B_e) e_i(k). \tag{16}$$

Let an integer s be sufficiently greater than n. Using the output equation (12b) and (16),

$$y_{k+i} = C\bar{A}^i x_k + C\mathcal{L}_i(\bar{A}, B) r_i(k) + C\mathcal{L}_i(\bar{A}, B_e) e_i(k) + e_{k+i} \tag{17}$$

is derived for $i = 0, \cdots, s-1$. Then, stack (17) for $i = 0, \cdots, s-1$, and we have the following equation:

$$y_s(k) = \mathcal{O}_y x_k + \mathcal{T}_y r_s(k) + \mathcal{H}_y e_s(k), \tag{18}$$

where $\mathcal{O}_y := \mathcal{O}_s(C, \bar{A})$, $\mathcal{T}_y := \mathcal{T}_s(\bar{A}, B, C, 0)$, $\mathcal{H}_y := \mathcal{T}_s(\bar{A}, B_e, C, I)$. Moreover, place (18) for $k = 0, \cdots, N$ side by side, and we obtain the following matrix IO equation:

$$\mathcal{Y}_f = \mathcal{O}_y \mathcal{X}_f + \mathcal{T}_y \mathcal{R}_f + \mathcal{H}_y \mathcal{E}_f \tag{19}$$

where define $\mathcal{X}_f := \mathcal{X}_{0,N}$ and the error matrix $\mathcal{E}_f := \mathcal{E}_{0,s,N}$.

Follow the same context discussed above, and we have the following equation with respect to $u_s(k)$:

$$u_s(k) = \mathcal{O}_u x_k + \mathcal{T}_u r_s(k) + \mathcal{H}_u e_s(k) \tag{20}$$

where $H := -KC$, and $\mathcal{O}_u := \mathcal{O}_s(H, \bar{A})$, $\mathcal{T}_u := \mathcal{T}_s(\bar{A}, B, H, 0)$, $\mathcal{H}_u := \mathcal{T}_s(\bar{A}, B_e, H, -K)$. Then, place (20) for $k = 0, \cdots, N$ side by side, and we derive the following matrix input-output equation:

$$\mathcal{U}_f = \mathcal{O}_u \mathcal{X}_f + \mathcal{T}_u \mathcal{R}_f + \mathcal{H}_u \mathcal{E}_f. \tag{21}$$

Now, recursive use of (16) toward the past gives the s-step ahead state equation between x_k and x_{k-s} as follows:

$$x_k = \bar{A}^s x_{k-s} + \mathcal{L}_r r_s(k-s) + \mathcal{L}_e e_s(k-s), \tag{22}$$

where $\mathcal{L}_r := \mathcal{L}_s(\bar{A}, B)$ and $\mathcal{L}_e := \mathcal{L}_s(\bar{A}, B_e)$. Then, place (22) for $k = 0, \cdots, N$ side by side, and we have the s-step ahead matrix state equation as follows:

$$\mathcal{X}_f = \bar{A}^s \mathcal{X}_p + \mathcal{L}_r \mathcal{R}_p + \mathcal{L}_e \mathcal{E}_p \tag{23}$$

where $\mathcal{X}_p := \mathcal{X}_{-s,N}$ and $\mathcal{E}_p := \mathcal{E}_{-s,s,N}$.

Finally, substitute (19) and (21) into (23) and notice an analogy between $\mathcal{U}_p$ and $\mathcal{U}_f$ with respect to (21), and we have the following matrix input-output equations:

$$\begin{bmatrix} \mathcal{U}_p \\ \mathcal{U}_f \\ \mathcal{Y}_f \end{bmatrix} = \begin{bmatrix} \mathcal{O}_u \\ \mathcal{O}_u\bar{A}^s \\ \mathcal{O}_y\bar{A}^s \end{bmatrix} \mathcal{X}_p + \begin{bmatrix} \mathcal{T}_u & 0 \\ \mathcal{O}_u\mathcal{L}_r & \mathcal{T}_u \\ \mathcal{O}_y\mathcal{L}_r & \mathcal{T}_y \end{bmatrix} \mathcal{R} + \begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u\mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y\mathcal{L}_e & \mathcal{H}_y \end{bmatrix} \mathcal{E}, \tag{24}$$

where $\mathcal{E} := \begin{bmatrix} \mathcal{E}_p^\mathsf{T} & \mathcal{E}_f^\mathsf{T} \end{bmatrix}^\mathsf{T}$.

L x k

With respect to the 3 block entries of the L-matrix, i.e., (11), the following lemma is derived from (8) directly:

Assume that the matrix $\mathcal{R}$ is of full row rank. Then, it holds that

$$\begin{bmatrix} L_{21} \\ L_{31} \\ L_{41} \end{bmatrix} = \begin{bmatrix} \mathcal{U}_p \\ \mathcal{U}_f \\ \mathcal{Y}_f \end{bmatrix} \mathcal{R}^\mathsf{T} L_{11}^{-\mathsf{T}} \tag{25}$$

P **f** From (8), we obtain

$$\begin{bmatrix} \mathcal{U}_p \\ \mathcal{U}_f \\ \mathcal{Y}_f \end{bmatrix} \mathcal{R}^\mathsf{T} = \begin{bmatrix} L_{21} \\ L_{31} \\ L_{41} \end{bmatrix} L_{11}^\mathsf{T}.$$

The invertibility of L_{11}^T is ensured by the assumption of the full row rankness of $\mathcal{R}$, and it completes the proof. □

Now, notice that $\mathcal{R}\mathcal{R}^\mathsf{T} = L_{11}L_{11}^\mathsf{T}$. Substitution of (24) into (25) yields

$$\begin{bmatrix} L_{21} \\ L_{31} \\ L_{41} \end{bmatrix} = \mathcal{S}_N + \mathcal{N}_N,$$

where $\mathcal{S}_N$ and $\mathcal{N}_N$ are called, respectively, the signal-based component and the noise-based component of (11), defined respectively as follows:

$$\mathcal{S}_N := \begin{bmatrix} \mathcal{O}_u \\ \mathcal{O}_u \bar{A}^s \\ \mathcal{O}_y \bar{A}^s \end{bmatrix} \mathcal{X}_p \mathcal{R}^\mathsf{T} L_{11}^{-\mathsf{T}} + \begin{bmatrix} \mathcal{T}_u & 0 \\ \mathcal{O}_u \mathcal{L}_r & \mathcal{T}_u \\ \mathcal{O}_y \mathcal{L}_r & \mathcal{T}_y \end{bmatrix} L_{11}, \quad (26)$$

$$\mathcal{N}_N := \begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix} \mathcal{E} \mathcal{R}^\mathsf{T} L_{11}^{-\mathsf{T}}. \quad (27)$$

The lower triangular matrix $\boldsymbol{L}_r$ is defined as the Cholesky factor, up to a sign matrix, of the following asymptotic covariance matrix:

$$\boldsymbol{L}_r \boldsymbol{L}_r^\mathsf{T} := \lim_{N\to\infty} \frac{1}{N} \mathcal{R}\mathcal{R}^\mathsf{T} = \text{block-diag}(\Omega_{rr}, \cdots, \Omega_{rr}).$$

For the rest of the paper, let us assume that the sign matrix is determined compatibly with the context.

The following theorem is for the probability convergence property of the signal-based component $\mathcal{S}_N$:

Under the assumptions from **1** to **6**,

$$\frac{1}{\sqrt{N}} \mathcal{S}_N \longrightarrow \mathcal{S}_\infty := \begin{bmatrix} \mathcal{T}_u & 0 \\ \mathcal{O}_u \mathcal{L}_r & \mathcal{T}_u \\ \mathcal{O}_y \mathcal{L}_r & \mathcal{T}_y \end{bmatrix} \boldsymbol{L}_r \quad \text{i.p. for } N \to \infty.$$

P f $\|\cdot\|_F$ denotes the Frobenius norm. For $\forall \varepsilon > 0$,

$$\begin{aligned} &\mathrm{P}\left[\left\| \frac{1}{\sqrt{N}} \mathcal{S}_N - \mathcal{S}_\infty \right\|_F > \varepsilon \right] \\ &\le \mathrm{P}\left[\left\| \begin{bmatrix} \mathcal{O}_u \\ \mathcal{O}_u \bar{A}^s \\ \mathcal{O}_y \bar{A}^s \end{bmatrix} \frac{1}{N} \mathcal{X}_p \mathcal{R}^\mathsf{T} \cdot \left(\frac{1}{\sqrt{N}} L_{11} \right)^{-\mathsf{T}} \right\|_F \right. \\ &\quad \left. + \left\| \begin{bmatrix} \mathcal{T}_u & 0 \\ \mathcal{O}_u \mathcal{L}_r & \mathcal{T}_u \\ \mathcal{O}_y \mathcal{L}_r & \mathcal{T}_y \end{bmatrix} \left(\frac{1}{\sqrt{N}} L_{11} - \boldsymbol{L}_r \right) \right\|_F > \varepsilon \right] \\ &\longrightarrow 0, \quad \text{as } N \to \infty. \end{aligned}$$

Note that $\lim_{N\to\infty} \frac{1}{N} \mathcal{X}_p \mathcal{R}^\mathsf{T} = 0$ holds from **5**. □

The following theorem is for the probability convergence property of the noise-based component $\mathcal{N}_N$:

Under the assumptions from **1** to **6**,

$$\frac{1}{\sqrt{N}} \mathcal{N}_N \longrightarrow 0 \quad \text{i.p. for } N \to \infty.$$

P f The assumption **4** gives $\lim_{N\to\infty} \frac{1}{N} \mathcal{E}\mathcal{R}^\mathsf{T} = 0$ Then, for $\forall \varepsilon > 0$, it holds that

$$\lim_{N\to 0} \mathrm{P}\left[\left\| \begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix} \left(\frac{1}{N} \mathcal{E}\mathcal{R}^\mathsf{T} \right) \cdot \left(\frac{1}{\sqrt{N}} L_{11} \right)^{-\mathsf{T}} \right\|_F \ge \varepsilon \right] = 0$$

□

The vectorization of the noise-based component yields

$$\mathrm{vec}\,(\mathcal{N}_N) = \left(Q_1^\mathsf{T} \otimes \begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix} \right) \mathrm{vec}\,(\mathcal{E}).$$

We introduce the notations: $\Phi(i) := \lim_{N\to\infty} Q_1^\mathsf{T} J_N^i Q_1$,

$$I_\nu := \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{bmatrix} \in \mathbb{R}^{\nu\times\nu}, \quad J_\nu := \begin{bmatrix} 0 & & & 0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix} \in \mathbb{R}^{\nu\times\nu},$$

The following theorem is for the asymptotic covariance matrix of $\mathrm{vec}(\mathcal{N}_N)$:

Under the assumptions from **1** to **6**, $\mathrm{vec}\,(\mathcal{N}_N)$ converges in distribution to the Gaussian distribution with zero mean and the covariance given by

$$\begin{aligned} &\lim_{N\to\infty} \mathrm{E}\left[\mathrm{vec}\,(\mathcal{N}_N)\, \mathrm{vec}\,(\mathcal{N}_N)^\mathsf{T} \right] \\ &= I_{2ms} \otimes \left(\begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix} (I_{2s} \otimes \Omega_{ee}) \begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix}^\mathsf{T} \right) \\ &\quad + \sum_{i=1}^{2s-1} \Phi(i) \otimes \left(\begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix} \left(\left(J_{2s}^\mathsf{T} \right)^i \otimes \Omega_{ee} \right) \begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix}^\mathsf{T} \right) \\ &\quad + \sum_{i=1}^{2s-1} \Phi(i)^\mathsf{T} \otimes \left(\begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix} \left(J_{2s}^i \otimes \Omega_{ee} \right) \begin{bmatrix} \mathcal{H}_u & 0 \\ \mathcal{O}_u \mathcal{L}_e & \mathcal{H}_u \\ \mathcal{O}_y \mathcal{L}_e & \mathcal{H}_y \end{bmatrix}^\mathsf{T} \right) \end{aligned}$$

P f The proof is omitted. See the literature (Oku and Ikeda, 2021). □

REFERENCES

Ikeda, K. (2014). On the precision of the plant estimates in some subspace identification methods. In *W*, 9516–9521. Cape Town.

Kojio, J., Ishibashi, H., Inoue, R., Ushida, S., and Oku, H. (2014). MIMO Closed-loop Subspace Model Identification and Hovering Control of a 6-DOF Coaxial Miniature Helicopter. In , 1679–1684. Sapporo.

Matsuba, I., Ushida, S., and Oku, H. (2012). MIMO Cosed-Loop Subspace Model Identification and Hovering Control of a Coaxial Mini Helicopter with 3 DOFs. In . Brussels, Belgium.

Nakayama, M. and Oku, H. (2018). An Experiment on Closed-loop System Identification of UAV Using Dual-rate Sampling. In , 598–603. Stockholm.

Oku, H. (2010). On asymptotic properties of MOESP-type closed-loop subspace model identification. In , 1015–1022. Kyoto, Japan.

Oku, H., Ogura, Y., and Fujii, T. (2006a). Closed-Loop Subspace Model Identification Based on QR-Factorization. In , 385–390. Kyoto, Japan.

Oku, H., Ogura, Y., and Fujii, T. (2006b). MOESP-type Closed-loop Subspace Model Identification Method. , 42(6), 636–642.

Oku, H. and Ikeda, K. (2021). On error analysis of a closed-loop subspace model identification method. , 54(9), 701–706.

Oku, H. and Ushida, S. (2009). Experiment on closed-loop subspace model identification of an unstable underactuated system. In , 4902–4907. Fukuoka.

Verhaegen, M. and Dewilde, P. (1992). Subspace model identification Part I: The output-error state space model identification class of algorithms. , 56(5), 1187–1210.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Small-gain conditions for robust stability of nonlinear infinite networks

Andrii Mironchenko * **Jochen Glück** **

* *Faculty of Computer Science and Mathematics, University of Passau, Innstraße 33, 94032 Passau, Germany (e-mail: andrii.mironchenko@uni-passau.de).*
** *Faculty of Mathematics and Natural Sciences, University of Wuppertal, Gaußstraße 33, 42119 Wuppertal, Germany (e-mail: glueck@uni-wuppertal.de).*

Abstract: We prove that a network of input-to-state stable infinite-dimensional systems is input-to-state stable, provided that the gain operator of the network satisfies the monotone limit property. This property is equivalent to the strong small-gain condition in the case of finite networks. We prove our small-gain theorem for a very general class of networks, including networks of nonlinear partial and delay differential equations. It also recovers the classical nonlinear small-gain theorems for finitely many finite-dimensional systems as a special case.

Keywords: large-scale systems, nonlinear control systems, small-gain theorems, input-to-state stability, distributed parameter systems.

1. INTRODUCTION

Stability of large-scale and infinite networks has attracted a lot of attention during the last decades.

A prominent place is occupied by the theory of linear spatially invariant systems Bamieh et al. (2002); Bamieh and Voulgaris (2005); Besselink and Johansson (2017); Curtain et al. (2009) whose applications range from microelectromechanical systems to control of car platoons. In these works, it is assumed that an infinite number of control systems are connected with each other by means of the same pattern. This nice geometrical structure, together with the linearity of subsystems, allowed the development of powerful criteria for the stability of infinite interconnections.

For finite networks consisting of nonlinear systems, groundbreaking results have been obtained within the framework of input-to-state stability (ISS), initiated for finite-dimensional systems in Sontag (1989), see also Sontag (2008), and recently extended to the infinite-dimensional setting, see Karafyllis and Krstic (2019); Mironchenko and Prieur (2020) for an overview of this topic. In this approach, the influence of any subsystem on other subsystems of a network is characterized by so-called gain functions. The gain operator constructed from these functions characterizes the interconnection structure of the network. The small-gain theorems for couplings of $n \in \mathbb{N}$ input-to-state stable systems of ordinary differential equations (ODEs) Jiang et al. (1994, 1996); Dashkovskiy et al. (2007, 2010) state that if the gains are small enough (i.e., the gain operator satisfies the small-gain condition), the network is stable. These results have been extended to finite networks of time-delay, and PDE systems Polushin et al. (2006); Tiwari et al. (2012); Dashkovskiy and Mironchenko (2013).

Very recently, a significant effort has been devoted to the development of small-gain theorems for infinite networks of ISS systems. In Dashkovskiy and Pavlichkov (2020) nonlinear small-gain theorems have been obtained under the assumption that all the gains are uniformly less than identity, which is a very strong assumption. In Kawan et al. (2021), tight small-gain theorems have been obtained for the networks of exponentially ISS systems with linear gains, while in Dashkovskiy et al. (2019), it is shown that ISS of a network can be guaranteed provided that there is a linear path of strict decay for the gain operator.

In this work, which is a conference version of the journal paper Mironchenko et al. (2021), we provide nonlinear ISS small-gain theorems for infinite networks with infinite-dimensional components. We do not impose any linearity and/or contractivity assumption for the gains, which makes the result truly nonlinear. Moreover, our result applies to networks of ODE systems, delay systems, PDE networks, and well-posed multi-physics systems. We derive our small-gain theorems in the trajectory formulation, in contrast to the papers Dashkovskiy et al. (2019); Dashkovskiy and Pavlichkov (2020); Kawan et al. (2021), where they have been shown in the Lyapunov formulation.

More precisely, we show that an infinite network consisting of ISS systems is ISS provided that the discrete-time system induced by the gain operator has the so-called *monotone limit property*. This property concerns the input-to-state behavior of the discrete-time control system $x(k+1) \leq \Gamma(x(k))+u(k)$ defined via the gain operator Γ on the positive cone of the gain space. For finite networks, this property is equivalent to the strong small-gain condition used in Dashkovskiy et al. (2007, 2010). Hence, our result fully covers the classical nonlinear small-gain theorems developed in Dashkovskiy et al. (2007, 2010).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

In this note, we omit all proofs and only present the main result. For much more on this topic, we refer the reader to the journal version of this paper Mironchenko et al. (2021). In the same paper, the relationships of the monotone limit property to various types of small-gain conditions have been studied, and a UGS small-gain theorem has been provided. For finitely many systems, our small-gain result has been shown in Mironchenko (2021), and we refer to this paper for the detailed discussion of the relationships between our main result and the available small-gain theorems for various classes of control systems, as well as for the other versions of small-gain theorems (e.g., small-gain theorems in maximum formulations).

Notation. We write $\mathbb{R}$ for the real numbers, $\mathbb{Z}$ for the integers, and $\mathbb{N}$ for natural numbers $1, 2, \ldots$. $\mathbb{R}_+$ and $\mathbb{Z}_+$ denote the sets of nonnegative reals and integers, respectively.

For a set U, we let $U^{\mathbb{R}_+}$ denote the space of all maps from $\mathbb{R}_+$ to U. By $\|w\|_{[0,t]}$ we denote the sup-norm of a bounded function $w : [0,t] \to W$, i.e., $\|w\|_{[0,t]} = \sup_{s\in[0,t]} \|w(s)\|_W$. Given a nonempty set I, we write $\ell_\infty(I)$ for the Banach space of all $x \in \mathbb{R}^I$ with $\|x\|_{\ell_\infty(I)} := \sup_{i\in I} |x(i)| < \infty$. Moreover, $\ell_\infty(I)^+ := \{x \in \ell_\infty(I) : x(i) \geq 0 \text{ for all } i \in I\}$. If $I = \mathbb{N}$, we simply write ℓ_∞ and ℓ_∞^+, respectively.

Throughout the paper, all considered vector spaces are vector spaces over $\mathbb{R}$. We use the standard classes $\mathcal{K}$, $\mathcal{K}_\infty$, $\mathcal{L}$, $\mathcal{KL}$ of comparison functions, see Kellett (2014).

2. CONTROL SYSTEMS AND THEIR STABILITY

In this paper, we define systems as follows.

Definition 2.1. Consider a triple $\Sigma = (X, \mathcal{U}, \phi)$ consisting of the following:

(i) A normed vector space $(X, \|\cdot\|_X)$, called the *state space*.

(ii) A vector space U of *input values* and a normed vector *space of inputs* $(\mathcal{U}, \|\cdot\|_\mathcal{U})$, where $\mathcal{U}$ is a linear subspace of $U^{\mathbb{R}_+}$. We assume that the following axioms hold:
- *The axiom of shift invariance*: for all $u \in \mathcal{U}$ and all $\tau \geq 0$, the time-shifted function $u(\cdot + \tau)$ belongs to $\mathcal{U}$ with $\|u\|_\mathcal{U} \geq \|u(\cdot + \tau)\|_\mathcal{U}$.
- *The axiom of concatenation*: for all $u_1, u_2 \in \mathcal{U}$ and for all $t > 0$ the concatenation of u_1 and u_2 at time t, defined by
$$u_1 \underset{t}{\diamondsuit} u_2(\tau) := \begin{cases} u_1(\tau) & \text{if } \tau \in [0,t], \\ u_2(\tau - t) & \text{otherwise} \end{cases}$$
belongs to $\mathcal{U}$.

(iii) A map $\phi : D_\phi \to X$, $D_\phi \subseteq \mathbb{R}_+ \times X \times \mathcal{U}$, called *transition map*, so that for all $(x,u) \in X \times \mathcal{U}$ it holds that $D_\phi \cap (\mathbb{R}_+ \times \{(x,u)\}) = [0, t_m) \times \{(x,u)\}$, for a certain $t_m = t_m(x,u) \in (0, +\infty]$. The corresponding interval $[0, t_m)$ is called the *maximal domain of definition* of the mapping $t \mapsto \phi(t,x,u)$, which we call a *trajectory* of the system.

The triple Σ is called a *(control) system* if it satisfies the following axioms:

(Σ1) *The identity property:* for all $(x,u) \in X \times \mathcal{U}$, it holds that $\phi(0,x,u) = x$.

(Σ2) *Causality:* for all $(t,x,u) \in D_\phi$ and $\tilde{u} \in \mathcal{U}$ such that $u(s) = \tilde{u}(s)$ for all $s \in [0,t]$, it holds that $[0,t] \times \{(x,\tilde{u})\} \subset D_\phi$ and $\phi(t,x,u) = \phi(t,x,\tilde{u})$.

(Σ3) *Continuity:* for each $(x,u) \in X \times \mathcal{U}$, the trajectory $t \mapsto \phi(t,x,u)$ is continuous on its maximal domain of definition.

(Σ4) *The cocycle property:* for all $x \in X$, $u \in \mathcal{U}$ and $t, h \geq 0$ so that $[0, t+h] \times \{(x,u)\} \subset D_\phi$, we have $\phi(h, \phi(t,x,u), u(t+\cdot)) = \phi(t+h, x, u)$.

We call the Σ *forward complete* if $D_\phi = \mathbb{R}_+ \times X \times \mathcal{U}$, i.e., $\phi(t,x,u)$ is defined for all $(t,x,u) \in \mathbb{R}_+ \times X \times \mathcal{U}$.

This class of systems encompasses control systems generated by ordinary differential equations, switched systems, time-delay systems, many classes of PDEs, important classes of boundary control systems, etc.

For the prolongation of trajectories of control systems, the following notion is of importance (Karafyllis and Jiang, 2011, Def. 1.4).

Definition 2.2. We say that a system Σ satisfies the *boundedness-implies-continuation (BIC) property* if for each $(x,u) \in X \times \mathcal{U}$ such that the maximal existence time $t_m = t_m(x,u)$ is finite, for any given $M > 0$ there exists $t \in [0, t_m)$ with $\|\phi(t,x,u)\|_X > M$.

Next, we introduce the input-to-state stability property, which unifies the classical asymptotic stability concept with the input-output stability notion, and is one of the cornerstones of nonlinear control theory as argued in Kokotović and Arcak (2001); Sontag (2008).

Definition 2.3. A system $\Sigma = (X, \mathcal{U}, \phi)$ is called *(uniformly) input-to-state stable (ISS)* if there exist $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}_\infty$ such that for all $(t,x,u) \in D_\phi$
$$\|\phi(t,x,u)\|_X \leq \beta(\|x\|_X, t) + \gamma(\|u\|_\mathcal{U}).$$

3. INFINITE INTERCONNECTIONS

This section introduces (feedback) interconnections of an arbitrary number of control systems, indexed by some nonempty set I. For each $i \in I$, let $(X_i, \|\cdot\|_{X_i})$ be a normed vector space which serves as the state space of a control system Σ_i. Before we can specify the space of inputs for Σ_i, we first have to construct the overall state space. Below, we use the sequence notation $(x_i)_{i\in I}$ for functions with domain I. The overall state space is then defined as
$$X := \left\{(x_i)_{i\in I} : x_i \in X_i,\ \forall i \in I \text{ and } \sup_{i\in I} \|x_i\|_{X_i} < \infty\right\}$$
and becomes a normed vector space with the norm
$$\|x\|_X := \sup_{i\in I} \|x_i\|_{X_i}.$$

We also define for each $i \in I$ the normed vector space $X_{\neq i}$ by the same construction as above, but for the restricted index set $I \setminus \{i\}$.

Now consider for each $i \in I$ a control system of the form
$$\Sigma_i = (X_i, \mathrm{PC}_b(\mathbb{R}_+, X_{\neq i}) \times \mathcal{U}, \bar{\phi}_i),$$
where $\mathrm{PC}_b(\mathbb{R}_+, X_{\neq i})$ is the space of all globally bounded piecewise continuous functions $w : \mathbb{R}_+ \to X_{\neq i}$, with the norm $\|w\|_\infty = \sup_{t\geq 0} \|w(t)\|_{X_{\neq i}}$. The norm on $\mathrm{PC}_b(\mathbb{R}_+, X_{\neq i}) \times \mathcal{U}$ is defined by

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$$\|(w,u)\|_{\mathrm{PC}_b(\mathbb{R}_+,X_{\neq i})\times\mathcal{U}} := \max\{\|w\|_\infty, \|u\|_\mathcal{U}\}. \quad (1)$$

Here we assume that $\mathcal{U} \subset U^{\mathbb{R}_+}$ for some vector space U, and $\mathcal{U}$ satisfies the axioms of shift invariance and concatenation. Then, by the definition of $\mathrm{PC}_b(\mathbb{R}_+, X_{\neq i})$ and the norm (1), these axioms are also satisfied for the product space $\mathrm{PC}_b(\mathbb{R}_+, X_{\neq i}) \times \mathcal{U}$.

Definition 3.1. Given the control systems Σ_i ($i \in I$) as above, we call a control system of the form $\Sigma = (X, \mathcal{U}, \phi)$ a *(feedback) interconnection* of the systems Σ_i if the following holds:

(i) The components ϕ_i of the transition map $\phi : D_\phi \to X$ satisfy
$$\phi_i(t,x,u) = \bar{\phi}_i(t, x_i, (\phi_{\neq i}, u)) \quad \text{for all } (t,x,u) \in D_\phi,$$
where $\phi_{\neq i}(\cdot) = (\phi_j(\cdot, x, u))_{j\in I\setminus\{i\}}$ for all $i \in I$. [1]

(ii) Σ has the BIC property.

We then call $X_{\neq i}$ the space of *internal input values*, $\mathrm{PC}_b(\mathbb{R}_+, X_{\neq i})$ the space of *internal inputs*, and $\mathcal{U}$ the space of *external inputs* of the system Σ_i. Moreover, we call Σ_i the *i-th subsystem* of Σ.

Let us define the concept of input-to-state stability for subsystems of a network.

Definition 3.2. Given the spaces $(X_j, \|\cdot\|_{X_j})$, $j \in I$, and the system Σ_i for a fixed $i \in I$, we say that Σ_i is *input-to-state stable (ISS) (in semimaximum formulation)* if Σ_i is forward complete and there are $\gamma_{ij}, \gamma_j \in \mathcal{K} \cup \{0\}$ for all $j \in I$, and $\beta_i \in \mathcal{KL}$ such that for all initial states $x_i \in X_i$, all internal inputs $w_{\neq i} = (w_j)_{j\in I\setminus\{i\}} \in \mathrm{PC}_b(\mathbb{R}_+, X_{\neq i})$, all external inputs $u \in \mathcal{U}$ and $t \geq 0$ it holds that
$$\|\bar{\phi}_i(t,x_i,(w_{\neq i},u))\|_{X_i} \leq \beta_i(\|x_i\|_{X_i}, t) + \sup_{j\in I}\gamma_{ij}(\|w_j\|_{[0,t]}) + \gamma_i(\|u\|_\mathcal{U}).$$

Here we assume that the functions γ_{ij} satisfy $\sup_{j\in I}\gamma_{ij}(r) < \infty$ for every $r \geq 0$ (implying that rhs is finite) and $\gamma_{ii} = 0$.

The functions γ_{ij} and γ_i in this definition are called *(nonlinear) gains*. For notational simplicity, we allow the case $\gamma_{ij} = 0$ for $j \neq i$.

Assuming that all systems Σ_i, $i \in I$, are ISS in semimaximum formulation, define a nonlinear monotone operator $\Gamma_\otimes : \ell_\infty(I)^+ \to \ell_\infty(I)^+$ from the gains γ_{ij} as follows:
$$\Gamma_\otimes(s) := \big(\sup_{j\in I}\gamma_{ij}(s_j)\big)_{i\in I}, \quad s = (s_i)_{i\in I} \in \ell_\infty(I)^+. \quad (2)$$

$\Gamma_\otimes$ is well-defined iff the following assumption holds.

Assumption 3.1. For every $r > 0$, we have
$$\sup_{i,j\in I}\gamma_{ij}(r) < \infty.$$

Also, observe that $\Gamma_\otimes$ is a monotone operator:
$$s^1 \leq s^2 \quad \Rightarrow \quad \Gamma_\otimes(s^1) \leq \Gamma_\otimes(s^2) \quad \text{for all } s^1, s^2 \in \ell_\infty(I)^+.$$

4. SMALL-GAIN THEOREM

We consider the system
$$x(k+1) \leq \Gamma_\otimes(x(k)) + u(k), \quad k \in \mathbb{Z}_+, \quad (3)$$
where $u \in \ell_\infty(\mathbb{Z}_+, \ell^+_\infty(I)) := \{u = (u(k))_{k\in\mathbb{Z}_+} : u(k) \in \ell^+_\infty(I),\ \|u\|_\infty := \sup_{k\in\mathbb{Z}_+}\|u(k)\|_X < \infty\}$.

As we will see, the stability properties of the interconnection will depend on the stability properties of the discrete-time system (3) induced by the gain operator.

Definition 4.1. System (3) has the *monotone limit property (MLIM)* if there is $\xi \in \mathcal{K}_\infty$ such that for every $\varepsilon > 0$ and $u \in \ell_\infty(\mathbb{Z}_+, \ell^+_\infty(I))$ and any solution $x(\cdot) = (x(k))_{k\in\mathbb{Z}_+}$ of (3) with $x(k+1) \leq x(k)$ for all $k \in \mathbb{Z}_+$, there exists $N = N(\varepsilon, u, x(\cdot)) \in \mathbb{Z}_+$ with
$$\|x(N)\|_X \leq \varepsilon + \xi(\|u\|_\infty).$$

Now we can state our main result.

Theorem 4.1. **(Nonlinear ISS small-gain theorem)** Let I be an arbitrary nonempty index set, $(X_i, \|\cdot\|_{X_i})$, $i \in I$ be normed spaces and $\Sigma_i = (X_i, \mathrm{PC}_b(\mathbb{R}_+, X_{\neq i}) \times \mathcal{U}, \bar{\phi}_i)$ be forward complete control systems. Assume that the interconnection $\Sigma = (X, \mathcal{U}, \phi)$ of the systems Σ_i is well-defined. Furthermore, let the following be satisfied:

(i) Each system Σ_i is ISS in the sense of Definition 3.2 with $\beta_i \in \mathcal{KL}$ and nonlinear gains $\gamma_{ij}, \gamma_i \in \mathcal{K} \cup \{0\}$.

(ii) There are $\beta_{\max} \in \mathcal{KL}$ and $\gamma_{\max} \in \mathcal{K}$ so that $\beta_i \leq \beta_{\max}$ and $\gamma_i \leq \gamma_{\max}$ pointwise for all $i \in I$.

(iii) Assumption 3.1 holds and the discrete-time system
$$w(k+1) \leq \Gamma_\otimes(w(k)) + v(k), \quad (4)$$
with $w(\cdot), v(\cdot)$ taking values in $\ell_\infty(I)^+$ has the MLIM property.

Then Σ is ISS.

If all interconnection gains γ_{ij} are linear, the small-gain condition in our theorem can be formulated more directly in terms of the gains, as the following corollary shows.

Corollary 4.1. **(Linear ISS small-gain theorem)** Given an interconnection $(\Sigma, \mathcal{U}, \phi)$ of systems Σ_i as in Theorem 4.1, additionally to the assumptions (i) and (ii) of this theorem, assume that all gains γ_{ij} are linear functions (and hence can be identified with nonnegative real numbers), $\Gamma_\otimes$ is well-defined and the following condition holds:
$$\lim_{n\to\infty}\Big(\sup_{j_1,\dots,j_{n+1}\in I}\gamma_{j_1j_2}\cdots\gamma_{j_nj_{n+1}}\Big)^{1/n} < 1. \quad (5)$$
Then Σ is ISS.

As an example, consider an infinite interconnection
$$\dot{x}_i = -x_i^3 + \max\{ax_{i-1}^3, bx_{i+1}^3, u\}, \quad i \in \mathbb{Z}, \quad (6)$$
where $a, b > 0$. Each Σ_i is a scalar system with the state $x_i \in \mathbb{R}$, internal inputs x_{i-1}, x_{i+1} and an external input u, belonging to the input space $\mathcal{U} := L_\infty(\mathbb{R}_+, \mathbb{R})$. Let the state space for the interconnection Σ be $X := \ell_\infty(\mathbb{Z})$.

It is not hard to show that this interconnection is well-posed. The stability of this network follows from:

Proposition 4.1. The coupled system (6) is ISS if and only if $\max\{a, b\} < 1$.

Proof. "⇒": For any $a, b > 0$ consider the scalar equation
$$\dot{z} = -(1 - \max\{a, b\})z^3,$$
subject to an initial condition $z(0) = x^*$. The function $y : t \mapsto (z(t))_{i\in\mathbb{Z}}$ is a solution of (6) subject to an initial

[1] By the causality axiom, we can assume that $\phi_{\neq i}$ is globally bounded, since $\bar{\phi}_i(t, x_i, (\phi_{\neq i}, u))$ does not depend on the values $\phi_{\neq i}(s)$ with $s > t$, and on the compact interval $[0, t]$, $\phi_{\neq i}$ is bounded because it is continuous.

condition $(x^*)_{i\in\mathbb{Z}}$ and input $u \equiv 0$. This shows that if $\max\{a, b\} \geq 1$, then the system (6) is not ISS.

"$\Leftarrow$": Consider x_{i-1}, x_{i+1} and u as inputs to the x_i-subsystem of (6) and define $q := \max\{ax_{i-1}^3, bx_{i+1}^3, u\}$. The derivative of $|x_i(\cdot)|$ along the trajectory satisfies for almost all t the following inequality:

$$\frac{\mathrm{d}}{\mathrm{d}t}|x_i(t)| \leq -|x_i(t)|^3 + q(t) \leq -|x_i(t)|^3 + \|q\|_\infty.$$

For any $\varepsilon > 0$, if $\|q\|_\infty \leq \frac{1}{1+\varepsilon}|x_i(t)|^3$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}|x_i(t)| \leq -\frac{\varepsilon}{1+\varepsilon}|x_i(t)|^3.$$

Arguing as in the proof of direct Lyapunov theorems ($x_i \mapsto |x_i|$ is an ISS Lyapunov function for the x_i-subsystem), see, e.g., (Sontag and Wang, 1995, Lem. 2.14), we obtain that there is $\beta \in \mathcal{KL}$ such that for all $t \geq 0$ it holds that

$$\begin{aligned}
&|x_i(t)| \leq \beta(|x_i(0)|, t) + \big((1+\varepsilon)\|q\|_\infty\big)^{\frac{1}{3}} \\
&= \beta(|x_i(0)|, t) + \max\{a_1\|x_{i-1}\|_\infty, b_1\|x_{i+1}\|_\infty, (1+\varepsilon)^{\frac{1}{3}}\|u\|_\infty^{\frac{1}{3}}\} \\
&\leq \beta(|x_i(0)|, t) + \max\{a_1\|x_{i-1}\|_\infty, b_1\|x_{i+1}\|_\infty\} + (1+\varepsilon)^{\frac{1}{3}}\|u\|_\infty^{\frac{1}{3}},
\end{aligned}$$

where $a_1 = (1+\varepsilon)^{\frac{1}{3}}a^{\frac{1}{3}}$, $b_1 = (1+\varepsilon)^{\frac{1}{3}}b^{\frac{1}{3}}$.

This shows that the x_i-subsystem is ISS in semimaximum formulation with the corresponding homogeneous gain operator $\Gamma : \ell_\infty^+(\mathbb{Z}) \to \ell_\infty^+(\mathbb{Z})$ given for all $s = (s_i)_{i\in\mathbb{Z}}$ by $\Gamma(s) = (\max\{a_1 s_{i-1}, b_1 s_{i+1}\})_{i\in\mathbb{Z}}$.

Previous computations are valid for all $\varepsilon > 0$. Now pick $\varepsilon > 0$ such that $a_1 < 1$ and $b_1 < 1$, which is possible as $a \in (0, 1)$ and $b \in (0, 1)$. The ISS of the network follows by Corollary 4.1. □

ACKNOWLEDGEMENTS

A. Mironchenko is supported by the German Research Foundation (DFG) via the grant MI 1886/2-2.

REFERENCES

Bamieh, B., Paganini, F., and Dahleh, M.A. (2002). Distributed control of spatially invariant systems. *IEEE Transactions on Automatic Control*, 47(7), 1091–1107.

Bamieh, B. and Voulgaris, P.G. (2005). A convex characterization of distributed control problems in spatially invariant systems with communication constraints. *Systems & Control Letters*, 54(6), 575–583.

Besselink, B. and Johansson, K.H. (2017). String stability and a delay-based spacing policy for vehicle platoons subject to disturbances. *IEEE Transactions on Automatic Control*, 62(9), 4376–4391.

Curtain, R., Iftime, O.V., and Zwart, H. (2009). System theoretic properties of a class of spatially invariant systems. *Automatica*, 45(7), 1619–1627.

Dashkovskiy, S. and Mironchenko, A. (2013). Input-to-state stability of infinite-dimensional control systems. *Mathematics of Control, Signals, and Systems*, 25(1), 1–35.

Dashkovskiy, S., Mironchenko, A., Schmid, J., and Wirth, F. (2019). Stability of infinitely many interconnected systems. *IFAC-PapersOnLine*, 52(16), 550–555.

Dashkovskiy, S. and Pavlichkov, S. (2020). Stability conditions for infinite networks of nonlinear systems and their application for stabilization. *Automatica*, 112, 108643.

Dashkovskiy, S., Rüffer, B., and Wirth, F. (2010). Small gain theorems for large scale systems and construction of ISS Lyapunov functions. *SIAM Journal on Control and Optimization*, 48(6), 4089–4118.

Dashkovskiy, S., Rüffer, B.S., and Wirth, F.R. (2007). An ISS small gain theorem for general networks. *Mathematics of Control, Signals, and Systems*, 19(2), 93–122.

Jiang, Z.P., Mareels, I.M.Y., and Wang, Y. (1996). A Lyapunov formulation of the nonlinear small-gain theorem for interconnected ISS systems. *Automatica*, 32(8), 1211–1215.

Jiang, Z.P., Teel, A.R., and Praly, L. (1994). Small-gain theorem for ISS systems and applications. *Mathematics of Control, Signals, and Systems*, 7(2), 95–120.

Karafyllis, I. and Jiang, Z.P. (2011). *Stability and Stabilization of Nonlinear Systems*. Springer, London.

Karafyllis, I. and Krstic, M. (2019). *Input-to-State Stability for PDEs*. Springer, Cham.

Kawan, C., Mironchenko, A., Swikir, A., Noroozi, N., and Zamani, M. (2021). A Lyapunov-based small-gain theorem for infinite networks. *IEEE Transactions on Automatic Control*, 66(12), 5830–5844.

Kellett, C.M. (2014). A compendium of comparison function results. *Mathematics of Control, Signals, and Systems*, 26(3), 339–374.

Kokotović, P. and Arcak, M. (2001). Constructive nonlinear control: A historical perspective. *Automatica*, 37(5), 637–662.

Mironchenko, A. (2021). Small gain theorems for general networks of heterogeneous infinite-dimensional systems. *SIAM Journal on Control and Optimization*, 59(2), 1393–1419.

Mironchenko, A., Kawan, C., and Glück, J. (2021). Nonlinear small-gain theorems for input-to-state stability of infinite interconnections. *Mathematics of Control, Signals, and Systems*, 33, 573–615.

Mironchenko, A. and Prieur, C. (2020). Input-to-state stability of infinite-dimensional systems: Recent results and open questions. *SIAM Review*, 62(3), 529–614.

Polushin, I.G., Tayebi, A., and Marquez, H.J. (2006). Control schemes for stable teleoperation with communication delay based on IOS small gain theorem. *Automatica*, 42(6), 905–915.

Sontag, E.D. (1989). Smooth stabilization implies coprime factorization. *IEEE Transactions on Automatic Control*, 34(4), 435–443.

Sontag, E.D. and Wang, Y. (1995). On characterizations of the input-to-state stability property. *Systems & Control Letters*, 24(5), 351–359.

Sontag, E.D. (2008). Input to state stability: Basic concepts and results. In *Nonlinear and Optimal Control Theory*, chapter 3, 163–220. Springer, Heidelberg.

Tiwari, S., Wang, Y., and Jiang, Z.P. (2012). Nonlinear small-gain theorems for large-scale time-delay systems. *Dynamics of Continuous, Discrete and Impulsive Systems Series A: Mathematical Analysis*, 19(1), 27–63.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Distributed control barrier function-based control scheme for multi-agent systems under a collective constraint ⋆

Xiao Tan * **Dimos V. Dimarogonas** *

* *School of EECS, KTH Royal Institute of Technology, Sweden*
(e-mail: xiaotan, dimos@kth.se*)*

Abstract: In this work, we consider multi-agent systems (MAS) operating under a collective constraint, i.e., a constraint that involves the collective states of MAS, using control barrier function (CBF) technique. CBF-based control design usually consists of designing a task-achieving controller, and modifying it minimally in a quadratic program (QP) to satisfy the state constraint. Despite its success for single-agent systems, most existing CBF-based control designs for MAS are either centralized or sub-optimal. Our proposed distributed CBF-based control scheme guarantees that the optimal to the QP control signals are obtained in finite time, and the collective constraint is satisfied for all time. The result is valid for a large class of MAS (linear or nonlinear, homogeneous or heterogeneous), underlying tasks (consensus, formation, coverage, etc), and collective constraints. We also analyze another scheme with some comparative remarks. Several numerical examples are shown.

Keywords: Multi-agent systems, constrained control, control barrier functions, distributed control, distributed optimization

1. INTRODUCTION

Control designs for dynamical systems under state/output constraints have been under extensive investigations in the literature. Many nonlinear control techniques are proposed, including potential fields, barrier-Lyapunov functions, and prescribed performance control. A recent technique called control barrier functions (CBF) (Xu et al. (2015); Tan et al. (2021)) has gained new attention. CBF provides a point-wise linear inequality condition on the system input, and by enforcing this condition at every state, the forward invariance of the safety set is guaranteed. This technique goes hand-in-hand with a computationally efficient, modular implementation leveraging quadratic programs that renders a pre-designed nominal controller to be safe in a minimal invasive manner. This methodology has been widely investigated and applied with practical success for single-agent systems.

Constraints considered for MAS are mainly defined locally in the literature, i.e., the satisfaction/violation of the constraints can be determined based on local information. These constraints are agent-wise and edge-wise state constraints; the later includes, e.g., collision avoidance, connectivity maintenance (see, e.g., Panagou et al. (2015)) and transient bounds of relative distances. In comparison, collective constraints are defined over the collective state of the MAS and cannot be evaluated to be violated or not using only local information. Some examples of collective constraints are the global connectivity maintenance (Capelli and Sabattini (2020)) and collective state boundedness constraint. Enforcing collective constraints in a distributed manner is in general more challenging since every agent can only obtain local information and make a decision based on it, while not knowing whether the constraint is satisfied. One common practice to enforce a collective constraint is to conservatively decompose it into several local ones.

There are many attempts extending the CBF framework to multi-agent systems. Rather straightforwardly, the CBF technique gives out a quadratic program with coupling constraints. In most works along this direction, the quadratic program is either solved in a centralized manner (Capelli and Sabattini (2020)), i.e, by a central module that has access to the states of every agent, or using a pre-allocation scheme (Wang et al. (2017)) that distributes the linear constraint among the agents involved. With the pre-allocation scheme, the optimality of the original quadratic program is generally lost. One could tackle this problem from a distributed optimization (Chen et al. (2020); Santilli et al. (2020)) perspective, which however no theoretical guarantees can be asserted regarding the satisfaction of the safety-certifying conditions during the solution iterations. This can potentially lead to unsafe behaviors. In general, a distributed implementation of CBF conditions is lacking.

In this extended abstract, we consider multi-agent systems under a collective constraint using control barrier function techniques. We propose two distributed implementations: one consists of local quadratic programs and an adaptive auxiliary variable that connects the local QPs. We show that, using this implementation scheme, the optimality

⋆ This work was supported by the Swedish Research Council (VR), the Swedish Foundation for Strategic Research (SSF), the ERC CoG LEAFHOUND, the EU CANOPIES project, and the Knut and Alice Wallenberg Foundation (KAW).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

condition of the original quadratic program will be fulfilled in finite time and the CBF condition is satisfied for all time. Main results have been reported in Tan and Dimarogonas (2021), yet some interesting but unmentioned aspects are identified here. Another implementation is more straightforward and leverages on average tracking algorithms, which, however, cannot guarantee the satisfaction of CBF condition for all time. Applications to a formation control problem with collective state boundedness constraint are demonstrated for both schemes.

2. PROBLEM FORMULATION

Consider a multi-agent system with N agents indexed by $\mathcal{I} = \{1, 2, ..., N\}$ whose communication graph $\mathcal{G} = (\mathcal{I}, E)$ is connected and undirected. $(i, j) \in E$ represents that the agents i, j can communicate with each other. The associated Laplacian matrix is denoted as L and the neighborhood set of agent i is defined as $N_i := \{j \in \mathcal{I} : (i, j) \in E\}$. The dynamics of agent $i \in \mathcal{I}$ is given by $\dot{\boldsymbol{x}}_i = \mathfrak{f}_i(\boldsymbol{x}_i) + \mathfrak{g}_i(\boldsymbol{x}_i)\boldsymbol{u}_i$, where the state $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$, and the control input $\boldsymbol{u}_i \in \mathbb{R}^{m_i}$, $\mathfrak{f}_i(\boldsymbol{x}_i)$, $\mathfrak{g}_i(\boldsymbol{x}_i)$ are locally Lipschitz functions in $\boldsymbol{x}_i$. We denote the stacked state $\boldsymbol{x} := (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_N^\top)^\top \in \mathbb{R}^n, n := \sum_{i\in\mathcal{I}} n_i$, the stacked control input $\boldsymbol{u} := (\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top, ..., \boldsymbol{u}_N^\top)^\top \in \mathbb{R}^m, m = \sum_{i\in\mathcal{I}} m_i$, the stacked vector fields $\mathfrak{f} = (\mathfrak{f}_1^\top, \mathfrak{f}_2^\top, ..., \mathfrak{f}_N^\top)^\top$ and $\mathfrak{g} = \text{blk}(\mathfrak{g}_1, \mathfrak{g}_2, ..., \mathfrak{g}_N)$. Thus, the stacked dynamics is

$$\dot{\boldsymbol{x}} = \mathfrak{f}(\boldsymbol{x}) + \mathfrak{g}(\boldsymbol{x})\boldsymbol{u}. \tag{1}$$

We denote the stacked locally available state to agent i as $\boldsymbol{x}_{loc,i} := (\boldsymbol{x}_i^\top, \boldsymbol{x}_{j_1}^\top, ..., \boldsymbol{x}_{j_{|N_i|}}^\top)^\top, j_k \in N_i$, for $k \in \{1, 2, ..., |N_i|\}$.

In this work, we consider collective constraints for MAS, i.e., the stacked state $\boldsymbol{x}$ is expected to evolve within a safety set, denoted as a superlevel set of a differentiable function $h : \mathbb{R}^n \to \mathbb{R}$:

$$\mathscr{C} = \{\boldsymbol{x} \in \mathbb{R}^n : h(\boldsymbol{x}) \geq 0\}. \tag{2}$$

Definition 1. (CBF). $h(\boldsymbol{x})$ in (2) is a control barrier function (CBF) for MAS (1) if there exists a locally Lipschitz extended class $\mathcal{K}$ function α such that:

$$\sup_{\boldsymbol{u}\in\mathbb{R}^m} [L_{\mathfrak{f}} h(\boldsymbol{x}) + L_{\mathfrak{g}} h(\boldsymbol{x})\boldsymbol{u} + \alpha(h(\boldsymbol{x}))] \geq 0, \ \forall \boldsymbol{x} \in \mathbb{R}^n \tag{3}$$

Xu et al. (2015); Tan et al. (2021) showed that any locally Lipschitz control input $\boldsymbol{u}$ that satisfies the CBF condition (3) renders the set $\mathscr{C}$ forward invariant and, if $\mathscr{C}$ is compact, asymptotically stable.

Assumption 1. The parameters in the CBF condition (3) are locally obtainable, i.e, the argument of (3) can be written in the form of

$$\sum_{i\in\mathcal{I}} \boldsymbol{a}_i^\top(\boldsymbol{x}_{loc,i})\boldsymbol{u}_i + \sum_{i\in\mathcal{I}} b_i(\boldsymbol{x}_{loc,i}) \leq 0. \tag{4}$$

This assumption admits a variety of collective constraints including, but not limited to, the examples below:

(1) $h(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} h_i(\boldsymbol{x}_i)$ with h_i differentiable. One example is the collective state boundedness constraint: $h(\boldsymbol{x}) = \sum_{i\in\mathcal{I}}(r^2 - \|\boldsymbol{x}_i\|^2)$ for some constant $r > 0$.
(2) $h(\boldsymbol{x}) = \sum_{l\in\mathscr{L}} h_l(\boldsymbol{x}_{l_i}, \boldsymbol{x}_{l_j})$, $\mathscr{L} \subset \mathbb{N}$ with $h_l(\cdot, \cdot)$ differentiable, $(l_i, l_j) \in E$. This could encode, for example, a least collective interaction level among all connected agents $h(\boldsymbol{x}) = \sum_{(i,j)\in E}(e^{-r_0\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2} - r_1)$ for some constants $0 < r_0, 0 < r_1 < 1$.

Remark 1. Assumption 1 indeed excludes some possible CBFs. However, the assumption could serve as a design principle for the communication topology of MAS if a particular collective constraint is given and, from Assumption 1, we know which communication links are relevant.

Assuming that nominal controllers are obtained by some distributed coordination protocol, i.e, $\boldsymbol{u}_{nom,i}(\boldsymbol{x}_{loc,i})$, we straightforwardly obtain the CBF-induced QP given as

$$\begin{aligned} &\min_{\boldsymbol{u}\in\mathbb{R}^m} \sum_{i\in\mathcal{I}} \frac{1}{2}\|\boldsymbol{u}_i - \boldsymbol{u}_{nom,i}(\boldsymbol{x}_{loc,i})\|^2 \\ &s.t. \ \sum_{i\in\mathcal{I}} \boldsymbol{a}_i^\top(\boldsymbol{x}_{loc,i})\boldsymbol{u}_i + \sum_{i\in\mathcal{I}} b_i(\boldsymbol{x}_{loc,i}) \leq 0. \end{aligned} \tag{5}$$

The intuition behind this QP is that the control signal is obtained by modifying the nominal controller subject to the safety condition in a minimum invasive manner. In the following we denote for brevity $\boldsymbol{a}_i, b_i, \boldsymbol{u}_{nom,i}$ when no ambiguity occurs. Defining $\bar{\boldsymbol{a}} = (\boldsymbol{a}_1^\top, \boldsymbol{a}_2^\top, ..., \boldsymbol{a}_N^\top)^\top$, $\bar{b} = \sum_{i\in\mathcal{N}} b_i$ and $\boldsymbol{u}_{nom} = (\boldsymbol{u}_{nom,1}^\top, \boldsymbol{u}_{nom,2}^\top, ..., \boldsymbol{u}_{nom,N}^\top)^\top$.

Remark 2. Here we assume agent dynamics to be control affine which encapsulates linear dynamics and a large class of nonlinear dynamics. Moreover, the agent dynamics and even the dimensions of state variables can be different among the agents. The distributed coordination protocol, which relates to the underlying tasks of the MAS, is designed independently.

Note that $\boldsymbol{a}_i, b_i, \boldsymbol{u}_{nom,i}$ are defined along the system trajectory, thus their values evolve with time. In the following we analyze the frozen-time optimality condition and provide two schemes that both guarantee the convergence to the time-varying solution optimal to QP in (5) in finite time.

3. MAIN RESULT

In this section, we will analyze the explicit solution to the QP in (5) and a distributed, yet equivalent QP, and then propose a distributed implementation that solves the original QP online while always enforcing the coupling constraint. We will also analyze another distributed implementation with some comparative remarks. Proofs of all the claims in Sec. 3.1 and Sec. 3.2 can be found in Tan and Dimarogonas (2021).

3.1 Explicit solution analysis and an equivalent QP

If the QP in (5) is feasible, then the optimal solution is explicitly given as

$$\boldsymbol{u}_i^\star = \boldsymbol{u}_{nom,i} - \max(0, (\bar{\boldsymbol{a}}^\top \boldsymbol{u}_{nom} + \bar{b})/\|\bar{\boldsymbol{a}}\|^2)\boldsymbol{a}_i, \forall i \in \mathcal{I}. \tag{6}$$

Although $\boldsymbol{u}_{nom,i}$ and $\boldsymbol{a}_i$ in (6) only require local information, the calculation of $(\bar{\boldsymbol{a}}^\top \boldsymbol{u}_{nom} + \bar{b})/\|\bar{\boldsymbol{a}}\|^2$ requires global information.

The QP problem in (5) can be equivalently given by

$$\begin{aligned} &\min_{(\boldsymbol{u},\boldsymbol{y})\in\mathbb{R}^{m+N}} \sum_{i\in\mathcal{I}} \frac{1}{2}\|\boldsymbol{u}_i - \boldsymbol{u}_{nom,i}\|^2 \\ &s.t. \ \boldsymbol{a}_i^\top \boldsymbol{u}_i + \sum_{j\in N_i} (y_i - y_j) + b_i \leq 0, \ \forall i \in \mathcal{I}, \end{aligned} \tag{7}$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_N) \in \mathbb{R}^N$ is an auxiliary decision variable. One could relate y_i with agent i and view $(\boldsymbol{x}_i, y_i)$ as an extended state of agent i. By analyzing its Lagrangian and KKT condition, we obtain, as expected, the optimal $\boldsymbol{u}_i^\star, i \in \mathcal{I}$, to (7) is the same as (6).

For the optimality condition for $\boldsymbol{y}^\star$, one observation is that $\boldsymbol{y}^\star$ is not unique: suppose $(\boldsymbol{u}^\star, \boldsymbol{y}')$ is an optimal solution to (7), then $(\boldsymbol{u}^\star, \boldsymbol{y}' + \beta \mathbf{1}_N), \beta \in \mathbb{R}$ is also an optimal solution. One sufficient condition on the optimal $\boldsymbol{y}^\star$ is

$$\boldsymbol{a}_i^\top \boldsymbol{u}_{nom,i} + \boldsymbol{l}_i \boldsymbol{y}^\star + b_i = c \boldsymbol{a}_i^\top \boldsymbol{a}_i, \ \forall i \in \mathcal{I} \tag{8}$$

with $c = (\bar{\boldsymbol{a}}^\top \boldsymbol{u}_{nom} + \bar{b})/\|\bar{\boldsymbol{a}}\|^2$.

3.2 Distributed Implementation

We propose a distributed implementation scheme that combines an adaptive law that locally updates y_i and a local QP with only the decision variable $\boldsymbol{u}_i$. Specifically, for each agent i, $\forall i \in \mathcal{I}$, we solve

$$\begin{aligned} &\min_{\boldsymbol{u}_i \in \mathbb{R}^{m_i}} \frac{1}{2}\|\boldsymbol{u}_i - \boldsymbol{u}_{nom,i}\|^2 \\ &s.t. \ \boldsymbol{a}_i^\top \boldsymbol{u}_i + \sum_{j \in N_i} (y_i - y_j) + b_i \le 0, \end{aligned} \tag{9}$$

and y_i is updated locally that will be derived later.

Proposition 1. If the local QPs given in (9) are feasible, i.e, $\sum_{j \in N_i}(y_i - y_j) + b_i \le 0$ whenever $\boldsymbol{a}_i = \mathbf{0}, \forall i \in \mathcal{I}$, then the solution $\bar{\boldsymbol{u}}^\star = (\boldsymbol{u}_1^\star, \boldsymbol{u}_2^\star, ..., \boldsymbol{u}_N^\star)$ to the local QPs satisfies $\bar{\boldsymbol{a}}^\top \bar{\boldsymbol{u}}^\star + \bar{b} \le 0$ for any value of $\boldsymbol{y}$.

This property is of interest because it states that whatever $\boldsymbol{y}$ is chosen, the safety guarantee is enforced whenever the local QPs are feasible.

Proposition 2. Assume that $\boldsymbol{a}_i \neq \mathbf{0}$ for all $i \in \mathcal{I}$. Define $c_i = \frac{1}{\boldsymbol{a}_i^\top \boldsymbol{a}_i}(\boldsymbol{l}_i \boldsymbol{y} + \boldsymbol{a}_i^\top \boldsymbol{u}_{nom,i} + b_i), i \in \mathcal{I}$ If $\boldsymbol{y}$ is chosen such that $c_i = c_j$ for any $i, j \in \mathcal{I}$, then the condition in (8) is satisfied.

In the following, an adaptive law for $\boldsymbol{y}$ is derived such that $\boldsymbol{c} = (c_1, c_2, ..., c_N)$ reaches a consensus in finite time even in the presence of slowly time-varying $\boldsymbol{a}_i, \boldsymbol{u}_{nom,i}, \boldsymbol{b}_i$. We note that the consensus algorithm is inspired by Franceschelli et al. (2014).

Proposition 3. Assume that $\boldsymbol{a}_i, \boldsymbol{u}_{nom,i}, \boldsymbol{b}_i$ are slowly time-varying in the sense that $\sum_{i \in \mathcal{I}} |\frac{d}{dt}(\boldsymbol{a}_i^\top \boldsymbol{a}_i) c_i(t) + \frac{d}{dt}(\boldsymbol{a}_i^\top \boldsymbol{u}_{nom,i} + b_i)| \le D$ for some $D > 0$ and $a_{min} \le \boldsymbol{a}_i^\top \boldsymbol{a}_i \le a_{max}$ for some positive constants a_{min}, a_{max} for all $i \in \mathcal{I}$. If the discontinuous adaptive law

$$\dot{\boldsymbol{y}} = -k_0 \text{sign}(L\boldsymbol{c}), \tag{10}$$

is applied, and the gain k_0 satisfies

$$k_0 \ge a_{max}(2\delta_{max} D / a_{min} + \epsilon), \tag{11}$$

where $\delta_{max} := \max_{i \in \mathcal{I}} |N_i|$, ϵ is a positive constant, then $c_i = c_j, \forall i, j \in \mathcal{I}$, within a finite time $t_r \le \frac{\|L\boldsymbol{c}(0)\|_1}{\epsilon}$.

Now we summarize the theoretical guarantees.

Theorem 1. Consider the CBF-induced quadratic program in (5). Assume that the conditions in Proposition 3 hold. Then the solution to the local QPs in (9) with $\boldsymbol{y}$ locally updated according to (10) solves the QP (5) in finite time. Moreover, the coupling constraint in (5) is satisfied for all time.

Remark 3. Instead of solving N local QPs as in (9) online, we could apply analytical solutions to the local QPs that are given by $\boldsymbol{u}_i = \boldsymbol{u}_{nom,i} - \max(0, c_i)\boldsymbol{a}_i, \ \forall i \in \mathcal{I}$.

3.3 Another distributed implementation

Taking a further look at the centralized optimal solution (6), we observe that the only variable that needs global information is $\frac{\sum_{i \in \mathcal{I}} \boldsymbol{a}_i^\top \boldsymbol{u}_{nom,i} + b_i}{\sum_{i \in \mathcal{I}} (\boldsymbol{a}_i^\top \boldsymbol{a}_i)}$. In this subsection, we propose another distributed implementation. For notational brevity, let $d_i(t) := \boldsymbol{a}_i^\top \boldsymbol{u}_{nom} + b_i$ and $e_i(t) := \boldsymbol{a}_i^\top \boldsymbol{a}_i$. Viewing $d_i(t)$ and $e_i(t)$ as two reference signals, we utilize distributed average tracking algorithms such that each agent tracks $\sum_{i \in \mathcal{I}} d_i/N$ and $\sum_{i \in \mathcal{I}} e_i/N$, respectively. Once each agent has the accurate averages, a local division between them gives $\sum_{i \in \mathcal{I}} d_i / \sum_{i \in \mathcal{I}} e_i$. More technical details are given below.

The scheme is implemented as follows. For agent i, define two extra local variables $p_i(t), q_i(t) \in \mathbb{R}$ whose dynamics are given by

$$\begin{aligned} \dot{z}_{1,i} &= \beta \sum_{j \in N_i} \text{sign}(p_j - p_i), z_{1,i}(0) = 0, \\ p_i(t) &= z_{1,i}(t) + d_i(t) \end{aligned} \tag{12}$$

and

$$\begin{aligned} \dot{z}_{2,i} &= \beta \sum_{j \in N_i} \text{sign}(q_j - q_i), z_{2,i}(0) = 0 \\ q_i(t) &= z_{2,i}(t) + e_i(t), \end{aligned} \tag{13}$$

where $z_{1,i}, z_{2,i}$ are local internal states. The input of agent i is chosen as

$$\boldsymbol{u}_i = \begin{cases} \boldsymbol{u}_{nom,i} - \max(0, p_i/q_i)\boldsymbol{a}_i, & \text{if } q_i \neq 0; \\ \boldsymbol{u}_{nom,i}, & \text{otherwise} \end{cases} \forall i \in \mathcal{I}. \tag{14}$$

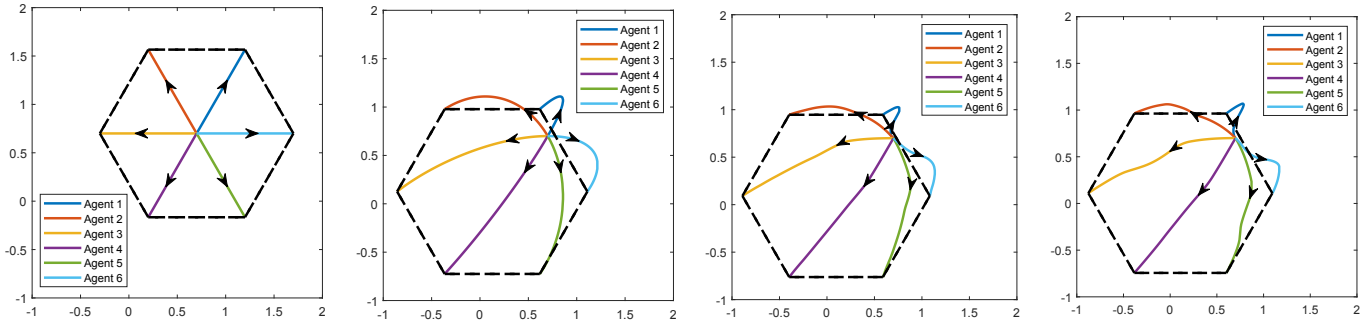
Theorem 2. Assume that $\sup_{t \in [0,\infty)} \max(|\dot{d}_i(t)|, |\dot{e}_i(t)|) \le \bar{d}$ for all $i \in \mathcal{I}$ and some positive constants $\bar{d}$, and $\sum_{i \in \mathcal{I}} e_i(t) > 0, \forall t > 0$. If we choose $\beta > \bar{d}$, then the agent input in (14) is optimal to the QP in (5) after time $T = \max\{\sum_{i \in \mathcal{I}} \sum_{j \in N_i} |p_j(0) - p_i(0)|, \sum_{i \in \mathcal{I}} \sum_{j \in N_i} |q_j(0) - q_i(0)|\}/2(\beta - \bar{d})$.

Proof. From Theorem 1 of Chen et al. (2012), since $\beta > \bar{d}$, $p_i(t), q_i(t)$ accurately tracks $\sum_{i \in \mathcal{I}} d_i/N$ and $\sum_{i \in \mathcal{I}} e_i/N$, respectively in finite time T. In view of $\boldsymbol{u}_i^\star$ given in (6), the input signal in (14) coincide with $\boldsymbol{u}_i^\star$ after T.

Remark 4. Both implementations assume slow dynamics of $\boldsymbol{a}_i, \boldsymbol{u}_{nom,i}$ and b_i, and use non-smooth analysis arguments to obtain a finite-time convergence result. One notable difference is that the implementation in Sec. 3.3 relaxes the requirement that $\boldsymbol{a}_i(t) \neq \mathbf{0}$ for all $i \in \mathcal{I}, t > 0$, which could be advantageous in many applications. However, it also fails to guarantee the satisfaction of the CBF condition during $[0, T]$, while the other one in Sec. 3.2 guarantees.

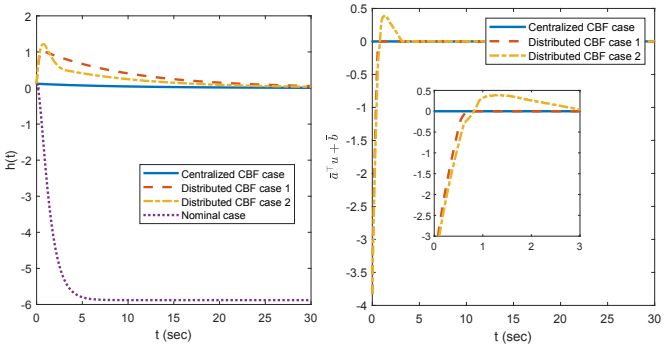
4. SIMULATIONS

In this section we demonstrate the efficacy of our proposed distributed schemes for a formation control problem under a collective state boundedness constraint. We consider 6 2-dimensional agents who has a simple single integrator dynamics. The MAS is tasked to form a star formation,



(a) Nominal case. (b) Centralized CBF case. (c) Distributed CBF case 1. (d) Distributed CBF case 2.

Fig. 1. MAS trajectories in four cases.



(a) CBF values. (b) CBF conditions.

Fig. 2. Time history of CBF values and CBF conditions.

i.e., the desired relative position $\boldsymbol{x}_d = (\boldsymbol{x}_{d,1}^\top, ..., \boldsymbol{x}_{d,6}^\top)^\top$ with $\boldsymbol{x}_{d,i} = (\cos(i\pi/3), \sin(i\pi/3)), i = 1, 2, ..., 6$. The communication graph is connected and undirected as shown in black dash lines in Fig. 1. All the agents start from $(0.7, 0.7)$ and the collective state is expected to evolve within the safety set $\{\boldsymbol{x} : h(\boldsymbol{x}) = 6 - \boldsymbol{x}^\top \boldsymbol{x} \geq 0\}$. Choose the extended class $\mathcal{K}$ function $\alpha(x) = 0.1x$. Straightforwardly, we obtain the CBF-induced QP in the form of (5) where $\boldsymbol{u}_{nom,i} = \sum_{i \in N_i}(\boldsymbol{x}_j - \boldsymbol{x}_{d,j} + \boldsymbol{x}_{d,i} - \boldsymbol{x}_i)$, $\boldsymbol{a}_i = 2\boldsymbol{x}_i, b_i = 0.1(\boldsymbol{x}_i^\top \boldsymbol{x}_i - 1)$.

We show the simulation results for the nominal case, where $\boldsymbol{u}_{nom,i}$ is used; the centralized CBF case, where $\boldsymbol{u}_i$ is obtained by solving the CBF-induced QP in a centralized way; the distributed CBF case 1, where $\boldsymbol{u}_i$ is obtained via the scheme in Sec. 3.2; and the distributed CBF case 2, where $\boldsymbol{u}_i$ is calculated as in Sec. 3.3. For a fair comparison, we choose both k_0 in (10) and β in (12) and (13) to be 1. The trajectories of the MAS are shown in Fig. 1. We observe that both distributed CBF cases have a very similar trajectory compared to that of the centralized CBF case. From Fig. 2(a), we also see that the collective boundedness constraint is fulfilled for all time in the centralized and distributed CBF cases. From Fig. 2(b), we observe that the CBF condition ($\bar{\boldsymbol{a}}^\top \boldsymbol{u} + \bar{b} \leq 0$) is always satisfied for the distributed scheme in Sec. 3.2, yet is temporally violated for the proposed scheme in Sec. 3.3.

5. CONCLUSION

In this work, we consider a CBF framework to control MAS under a collective constraint. In particular, two distributed implementation schemes to implement the CBF-induced quadratic programs are discussed. Under the assumption that the parameters of the coupling constraint are slowly time-varying, both proposed implementations solve the CBF-induced QP in finite time. One of them guarantees the satisfaction of the coupling constraint for all time, while the other one is applicable even when some coefficients of the agent inputs become zero.

REFERENCES

Capelli, B. and Sabattini, L. (2020). Connectivity maintenance: Global and optimized approach through control barrier functions. In *2020 IEEE Int. Conf. Robot. Autom. (ICRA)*, 5590–5596. IEEE.

Chen, F., Cao, Y., and Ren, W. (2012). Distributed average tracking of multiple time-varying reference signals with bounded derivatives. *IEEE Trans. Auto. Cont.*, 57(12), 3169–3174.

Chen, Y., Santillo, M., Jankovic, M., and Ames, A.D. (2020). Online decentralized decision making with inequality constraints: an ADMM approach. *IEEE Control Syst. Lett.*, 5(6), 2156–2161.

Franceschelli, M., Pisano, A., Giua, A., and Usai, E. (2014). Finite-time consensus with disturbance rejection by discontinuous local interactions in directed graphs. *IEEE Trans. Auto. Cont.*, 60(4), 1133–1138.

Panagou, D., Stipanović, D.M., and Voulgaris, P.G. (2015). Distributed coordination control for multi-robot networks using lyapunov-like barrier functions. *IEEE Trans. Auto. Cont.*, 61(3), 617–632.

Santilli, M., Oliva, G., and Gasparri, A. (2020). Distributed finite-time algorithm for a class of quadratic optimization problems with time-varying linear constraints. In *2020 IEEE Conf. Decis. Control (CDC)*, 4380–4386. IEEE.

Tan, X. and Dimarogonas, D.V. (2021). Distributed implementation of control barrier functions for multi-agent systems. *IEEE Control Syst. Lett.*, 6, 1879–1884.

Tan, X., Shaw Cortez, W., and Dimarogonas, D.V. (2021). High-order barrier functions: robustness, safety and performance-critical control. *IEEE Trans. Auto. Cont.*

Wang, L., Ames, A., and Egerstedt, M. (2017). Safety barrier certificates for collisions-free multirobot systems. *IEEE Trans. on Robot.*, 33(3), 661–674.

Xu, X., Tabuada, P., Grizzle, J.W., and Ames, A.D. (2015). Robustness of control barrier functions for safety critical control. In *Proc. IFAC Conf. Anal. Design Hybrid Syst.*, volume 48, 54–61.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Approximation Theory for Deep Learning in Time Series Analysis

Haotian Jiang * **Zhong Li** ** **Qianxiao Li** ***

* *Department of Mathematics, National University of Singapore, Singapore 119077 (e-mail: e0012663@u.nus.edu).*
** *School of Mathematical Sciences, Peking University, Beijing 100080 (e-mail: li_zhong@pku.edu.cn).*
*** *Department of Mathematics, National University of Singapore, Singapore 119077 (e-mail: qianxiao@nus.edu.sg).*

Abstract: This extended abstract summarizes a series of recent works (Li et al., 2021a,b, 2022; Jiang et al., 2021) on the approximations theory of deep learning methods for time series modelling and analysis. The primary aim is to develop the mathematical foundations for modelling sequential relationships with neural networks, which guides the practical implementation and design of such architectures. In particular, we place on concrete mathematical footing on when and how certain architectures (recurrent neural networks, encoder-decoder structures, dilated convolutions, etc) can adapt to corresponding data structures in the temporal relationships to be learned (memory, rank, sparsity, etc). These form the first step towards principled neural network architecture design and selection for practical machine learning of temporal relationships.

Keywords: Approximation theory, time series analysis, deep learning.

1. INTRODUCTION

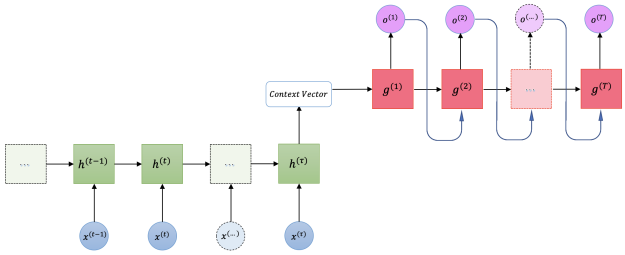


Fig. 1. A recurrent encoder-decoder architecture which maps a sequence into another sequence Cho et al. (2014).

The primary interest of this series of works is to analyze sequence to sequence (seq2seq) modelling, where the input and output are both time-series or sequences. Fig. 1 illustrates an example of a deep learning architecture for seq2seq modelling. There are various architectures that can achieve state-of-the-art performance in different tasks, e.g. the Transformers (Vaswani et al., 2017) in natural language processing (NLP) and the WaveNet (Oord et al., 2016) in audio signal processing, and Vision Transformers (Vit, Dosovitskiy et al. (2020)) in image classification. Our goal is to rigorously formulate the seq2seq problems and theoretically understand how these architectures fundamentally differ from each other. We developed a mathematical framework for analysing seq2seq problems, and this framework enables us to compare different architectures under the same regime. Currently, our work covers recurrent neural networks (RNNs), RNN encoder-decoders and convolutional neural networks (CNNs). Active research is also in process to include more architectures, including attention mechanism and Transformers.

This extended abstract is organised as follows. In Section 2, we will introduce the analysis framework including problem settings and notations. In the remaining sections, we will present and discuss the main results for different architectures.

2. GENERAL FORMULATION OF SEQ2SEQ MODELLING PROBLEM

2.1 Definition of sequences

We view a sequence $\boldsymbol{x} : \mathcal{I} \to \mathbb{R}^d$ as a function from a index set to real vectors. Based on specific settings, the index set $\mathcal{I}$ may be taken as (a subset of) $\mathbb{R}$ (continuous-time setting) or (a subset of) $\mathbb{Z}$ (discrete-time setting).

2.2 Formulation of temporal relationships

For seq2seq problems, we need to model mappings between two sequences. We denote the input and output space by $\mathcal{X}$ and $\mathcal{Y}$, respectively. The input space is usually taken as a normed vector space with some norm $\|\boldsymbol{x}\|_{\mathcal{X}}$. The mapping between $\boldsymbol{x}$ and $\boldsymbol{y}$ can be described by a sequence of functionals:

$$\{y(t) = H_t(\boldsymbol{x}) : t \in \mathcal{I}\}. \tag{1}$$

The output at each time depend on the entire input sequence through a time-dependent functional. The goal is to learn this sequence of functionals $\boldsymbol{H} := \{H_t : t \in \mathcal{I}\}$, which is called the target temporal relationship. The induced functional norm is defined as

$$\|H_t\| := \sup_{\boldsymbol{x}\in\mathcal{X},\|\boldsymbol{x}\|_{\mathcal{X}}\le 1} |H_t(\boldsymbol{x})|, \tag{2}$$

$$\|\boldsymbol{H}\| := \sup_{t\in\mathcal{I}} \|H_t\|. \tag{3}$$

We denote a temporal relationship built from a specific architecture/model by $\hat{\boldsymbol{H}}$. The hypothesis space and concept space is denoted by $\mathcal{H}$ and $\mathcal{C}$, respectively.

Here, we introduce an example to elaborate this functional formulation. The convolution can be considered as a seq2seq relation in the following sense. Suppose we have an input $\boldsymbol{x}$ and a convolution filter g, then the output at time t is given by $y_t = H_t(\boldsymbol{x}) = \int_{-\infty}^{\infty} g(s)x(t-s)ds$, hence the sequence of functionals $\boldsymbol{H} = \{H_t : t \in \mathbb{R}\}$ can be viewed as the entire output sequence $\boldsymbol{y}$.

We are interested in the approximation capabilities of different models, with the approximation error defined as

$$\|\boldsymbol{H} - \hat{\boldsymbol{H}}\|. \tag{4}$$

As is shown later, different models are good at approximating specific classes of target temporal relationships.

2.3 Properties of temporal relationships

We assume that the target temporal relationship (in a certain concept space $\mathcal{C}$) satisfy some specific properties. Below is a list of the related properties possibly used in our works.

P1. $H_t \in \boldsymbol{H}$ is a linear and continuous functional, if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$\begin{aligned} H_t(\lambda_1\boldsymbol{x}_1 + \lambda_2\boldsymbol{x}_2) &= \lambda_1 H_t(\boldsymbol{x}_1) + \lambda_2 H_t(\boldsymbol{x}_2), \\ \|H_t\| &= \sup_{\boldsymbol{x}\in\mathcal{X},\|\boldsymbol{x}\|_{\mathcal{X}}\le 1} |H_t(\boldsymbol{x})| < \infty, \end{aligned} \tag{5}$$

$\boldsymbol{H}$ is linear and continuous if each H_t is linear and continuous.

P2. $H_t \in \boldsymbol{H}$ is regular, if for any sequence $\{\boldsymbol{x}_n : n \in \mathbb{N}\}$ such that $x_n(s) \to 0$ for almost every $s \in \mathcal{I}$, we have $\lim_{n\to\infty} H_t(\boldsymbol{x}_n) = 0$. $\boldsymbol{H}$ is regular if each H_t is regular.

P3. $\boldsymbol{H}$ is causal, if it does not depend on the future inputs. That is, for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and any $t \in \mathcal{I}$ such that $x_1(s) = x_2(s)$ for all $s \le t$, the output satisfies $H_t(\boldsymbol{x}_1) = H_t(\boldsymbol{x}_2)$.

P4. A sequence of functionals $\boldsymbol{H}$ is time-homogeneous, if for any $t, \tau \in \mathcal{I}$, $H_t(\boldsymbol{x}) = H_{t+\tau}(\boldsymbol{x}^{(\tau)})$ with $x^{(\tau)}(s) := x(s-\tau)$ for all $s \in \mathcal{I}$.

3. RECURRENT NEURAL NETWORKS

In the continuous-time setting, the simplest RNN (with skip connections and one linear readout layer) is given by

$$\begin{aligned} \frac{d}{dt}h(t) &= \sigma(Wh(t) + Ux(t)), \\ \hat{y}(t) &= c^\top h(t), \\ \text{with } c \in \mathbb{R}^m, W \in \mathbb{R}^{m\times m}, U &\in \mathbb{R}^{m\times d}. \end{aligned} \tag{6}$$

Here, $h \in \mathbb{R}^m$ is the hidden state and m denotes the width of RNNs, which determines the model complexity. If we assume the linearity, (6) defines a family of functionals

$$\mathcal{H}^{(m)}_{\text{RNN}} := \Big\{\hat{\boldsymbol{H}} : \hat{H}_t = \hat{y}(t) = \int_0^\infty c^\top e^{Ws} U x(t-s) ds\Big\}. \tag{7}$$

The hypothesis space for RNNs with arbitrary widths is defined as

$$\mathcal{H}_{\text{RNN}} := \bigcup_{m\in\mathbb{N}_+} \mathcal{H}^{(m)}_{\text{RNN}}. \tag{8}$$

The concept space we considered here is

$$\mathcal{C} = \{\boldsymbol{H} : \boldsymbol{H} \text{ satisfies P1, P2, P3, P4}\}. \tag{9}$$

We next present the main result.

3.1 Approximation rate (Li et al., 2021a, Theorem 4.2)

Let $y_i(t) = H_t(\boldsymbol{e}_i)$, $i = 1, \dots, d$, where $\boldsymbol{e}_i$ is a step constant signal. Suppose

- $y_i \in C^{(\alpha+1)}(\mathbb{R})$;
- $e^{\beta t} y_i^{(k)}(t) = o(1)$.

We have the estimate

$$\|\boldsymbol{H} - \hat{\boldsymbol{H}}\| \le \frac{C(\alpha)\gamma d}{\beta m^\alpha}, \tag{10}$$

where α measures the smoothness of the target, and β measures the memory/decay of the target. $C(\alpha)$ and γ are two positive universal constants depending on α, β. The smoothness and memory are characterised by y_i, i.e. the response of target functionals under (step) constant inputs.

This result implies that a target temporal relationship can be efficiently approximated by RNNs if it is smooth (large α) and decays fast (large β). Conversely, the target with sudden changes are difficult to be learned by RNNs.

4. RNN ENCODER-DECODER

The architecture with RNNs as both encoder and decoder can be formulated as

$$\begin{aligned} h_s &= \sigma_E(W_E h_{s-1} + U_E x_s + b_E), \quad & v &= h_\tau, \\ g_t &= \sigma_D(W_D g_{t-1} + b_D), & g_0 &= v, \\ o_t &= W_O g_t + b_O, \end{aligned} \tag{11}$$

where h_t, g_t are hidden states of the encoder and decoder, respectively. We take a linear, residual and continuous-time idealisation of (11):

$$\begin{aligned} \frac{d}{ds}h_s &= Wh_s + Ux_s, \quad & v &= Qh_0, \quad & s \le 0, \\ \frac{d}{dt}g_t &= Vg_t, & g_0 &= Pv, \\ y_t &= c^\top g_t, & t &\ge 0, \\ \text{with } W \in \mathbb{R}^{m_E\times m_E}, V &\in \mathbb{R}^{m_D\times m_D}, U \in \mathbb{R}^{m_E\times d}, \\ Q \in \mathbb{R}^{N\times m_E}, P &\in \mathbb{R}^{m_D\times N}. \end{aligned} \tag{12}$$

Here, m_E and m_D denote the width of encoder and decoder, respectively. N is the dimension of the coding vector v, and it can be also understood as the model rank. (12) defines a family of functionals

$$\mathcal{H}^{(m_E,m_D,N)}_{\text{EncDec}} := \Big\{\hat{\boldsymbol{H}} : \hat{H}_t = \int_0^\infty c^\top e^{Vt} P Q e^{Ws} U x_{-s} ds\Big\}. \tag{13}$$

The hypothesis space for RNN encoder-decoders with arbitrary widths is defined by

$$\mathcal{H}_{\text{EncDec}} := \bigcup_{m_E,m_D,N\in\mathbb{N}_+} \mathcal{H}^{(m_E,m_D,N)}_{\text{EncDec}}. \tag{14}$$

The concept space we considered here is

$$\mathcal{C} = \{\boldsymbol{H} : \boldsymbol{H} \text{ satisfies P1, P2}\} \tag{15}$$

We next present the main result.

4.1 Approximation rate (Li et al., 2021b, Theorem 4.3)

Let $y_i(t,s) = H_t(e_i \mathbf{1}_{(-\infty,-s]}),\ i = 1,\dots,d$. Suppose

- $y_i \in C^{(\alpha+1)}([0,\infty))$;
- $e^{\beta(t+s)} \frac{\partial^{k+l}}{\partial t^k \partial s^l} y_i(t,s) = o(1)$ as $\|(t,s)\| \to \infty$;
- $\{\sigma_n\}$ are singular values of ρ. Intuitively, we have $\rho(t,s) = \sum_{n=1}^{\infty} \sigma_n \varphi_n(t) \phi_n(s)$, where $\{\varphi_n\}$ and $\{\phi_n\}$ are orthonormal bases, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ denote the singular values.

We have

$$\|\boldsymbol{H} - \hat{\boldsymbol{H}}\| \leq \frac{C(\alpha)\gamma d}{\beta^2} \Bigg\{ \left(1 + \sqrt{\bar{m} - N}\right) \cdot \left(\frac{1}{m_E^{\alpha}} + \frac{1}{m_D^{\alpha}}\right) + \left(\sum_{n=N+1}^{\bar{m}} \sigma_n^2\right)^{\frac{1}{2}} + \left(\sum_{n=N+1}^{\bar{m}} \sigma_n\right)^{\frac{1}{2}} \left(\frac{1}{m_E^{\alpha/2}} + \frac{1}{m_D^{\alpha/2}}\right) \Bigg\}, \tag{16}$$

where α measures the smoothness of the target, β measures the memory/decay of the target. $C(\alpha)$ and γ are two positive universal constants depending on α, β. A target can be efficiently approximated by RNN encoder-decoders if it is smooth (large α) and decays fast (large β). These are almost the same as RNNs.

New insights are presented as follows. Different from the RNN case where we only consider time-homogeneous target relationships, here we instead investigate time-inhomogeneous relationships. We have $\mathcal{H}_{\text{RNN}} \subseteq \mathcal{H}_{\text{EncDec}}$, which implies that the encoder-decoder is more general with the capability to learn time-inhomogeneity. In addition, the time-inhomogeneous temporal relationship considered here possesses a two parameter representation $\rho(t,s)$, leading to an extra characterisation which we called temporal product structure. Intuitively, we can decompose $\rho(t,s)$ along the input and output temporal direction, i.e. s and t, such that $\rho(t,s) = \sum_{n=1}^{\infty} \sigma_n \varphi_n(t) \phi_n(s)$, which is analogous to the singular value decomposition of matrices. We say $\rho(t,s)$ has a low "effective rank" if the singular values $\{\sigma_n\}$ decays fast. Our approximation rate shows that a target with low effective rank can be efficiently approximated by the encoder-decoder architecture with a small model rank N, which reduces the number of parameters needed to achieve the accuracy.

5. CONVOLUTIONAL NEURAL NETWORKS

For CNNs, we use the discrete-time setting. Inspired from the dilated convolution architecture in WaveNet (Oord et al., 2016), a dilated convolution-based temporal sequence model with K layers and M_k channels at layer k is given by

$$\begin{aligned} \boldsymbol{h}_{0,i} &= \boldsymbol{x}_i, \\ \boldsymbol{h}_{k+1,i} &= \sigma\left(\sum_{j=1}^{M_k} \boldsymbol{w}_{kji} *_{d_k} \boldsymbol{h}_{k,j}\right), \\ \hat{\boldsymbol{y}} &= \boldsymbol{h}_K. \end{aligned} \tag{17}$$

Here, $*_{d_k}$ denotes the dilated convolution operator with the dilation rate equalling to d_k, $\boldsymbol{x}_i$ is the i^{th} dimension of $\boldsymbol{x}$, and $\boldsymbol{w}_{kji}$ is the filter from channel j at layer k to channel i at layer $k+1$. All the filters have a size $l \geq 2$. If σ is linear, we have

$$\mathcal{H}_{\text{CNN}}^{(l,K,\{M_k\})} := \left\{ \hat{\boldsymbol{H}} : \hat{H}_t(\boldsymbol{x}) = \sum_{s \in \mathbb{N}} \rho^{(\hat{\boldsymbol{H}})}(s)^{\top} x(t-s) \right\}, \tag{18}$$

where $\rho^{(\hat{\boldsymbol{H}})}$ is a finite-supported vector. The hypothesis space for CNNs with arbitrary depths and number of channels is defined as

$$\mathcal{H}_{\text{CNN}}^{(l)} = \bigcup_{K \in \mathbb{N}_+} \bigcup_{\{M_k\} \in \mathbb{N}_+^K} \mathcal{H}_{\text{CNN}}^{(l,K,\{M_k\})}. \tag{19}$$

The concept space we considered here is

$$\mathcal{C} = \{\boldsymbol{H} : \boldsymbol{H} \text{ satisfies P1, P2, P3, P4}\}. \tag{20}$$

We next present the main result.

5.1 Approximation rate (Jiang et al., 2021, Theorem 4)

Define the complexity measure of $\boldsymbol{H}$ by

$$\begin{aligned} &\mathrm{C}^{(l,g)}(\boldsymbol{H}) \\ &= \inf\left\{ c : \left(\sum_{i=s+K}^{lK} |\sigma_i^{(K)}|^2\right)^{\frac{1}{2}} \leq c g(s),\ s \geq 0, K \in \mathbb{N}_+ \right\}, \end{aligned} \tag{21}$$

where $\sigma_1^{(K)} \geq \sigma_2^{(K)} \geq \cdots \geq \sigma_{lK}^{(K)} \geq 0$ denote the singular values of the tensorisation of $\boldsymbol{\rho}^{(\boldsymbol{H})}$, and g is a non-increasing function with zero limit at positive infinity. For any $\boldsymbol{H} \in \mathrm{C}^{(l,g)}$ and any set of parameters $(K, \{M_k\})$, we have

$$\|\boldsymbol{H} - \hat{\boldsymbol{H}}\| \leq d\ g(KM^{\frac{1}{K}} - K)\, \mathrm{C}^{(l,g)}(\boldsymbol{H}) + \|\boldsymbol{\rho}^{(\boldsymbol{H})}_{[l^K,\infty)}\|_2, \tag{22}$$

where $M := \frac{1}{d}(\sum_{k=2}^{K} M_k M_{k-1} - lK)$ denotes the "effective" number of filters.

In terms of approximation, CNNs are different from the former recurrent architectures. Our approximation rate shows that a target with low effective (tensor) rank can be easily approximated by CNNs with fewer channels $\{M_k\}$, which controls the model (tensor) rank. The target with sparsity is a special case, which is also low rank and easy to be approximated (since CNNs can construct each non-zero value individually).

Different from RNNs (possessing an infinite long filter), the CNN has a finite-supported filter (despite with exponentially large receptive fields). Hence, if the target has long tails, more parameters are required for CNNs to learn it.

6. DISCUSSION

In this section, we discuss the relations between different architectures. First, RNNs and RNN encoder-decoders are similar with each other, since both of them possess recurrent architectures. When an RNN encoder-decoder is required to be time-homogeneous, it becomes an RNN. In particular, we have $\mathcal{H}_{\text{RNN}} \subseteq \mathcal{H}_{\text{EncDec}}$. Similar with RNNs, it is also difficult for RNN encoder-decoders to handle targets with long-term memories or sudden changes.

In addition, RNNs and CNNs are both causal and time-homogeneous, but they have different underlying structures. The RNN uses a power sum filter, which is good at approximating targets with fast decayed structures, but not efficient to handle sudden changes or long-term memories. While the CNN uses a finite-supported filter, which works well for targets with low (effective) ranks or fast decayed singular values. The approximation can become inefficient if the tail error is significant, or the finite-supported part does not possess this low (effective) rank structure.

Finally, we want to clarify that the rank concept appeared in CNNs and encoder-decoders are totally different. In the CNN setting, the rank corresponds to tensors after certain tensorisation. While for encoder-decoders, the rank is related with the temporal product structure.

ACKNOWLEDGEMENTS

HJ is supported by National University of Singapore under PGF scholarship. ZL is supported by Peking University under BICMR mathematical scholarship. QL is supported by the National Research Foundation, Singapore, under the NRF fellowship (NRF-NRFF13-2021-0005).

REFERENCES

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. URL http://arxiv.org/abs/1406.1078. ArXiv: 1406.1078.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. URL https://openreview.net/forum?id=YicbFdNTTy.

Jiang, H., Li, Z., and Li, Q. (2021). Approximation Theory of Convolutional Architectures for Time Series Modelling. In *Proceedings of the 38th International Conference on Machine Learning*, 4961–4970. PMLR. URL https://proceedings.mlr.press/v139/jiang21d.html. ISSN: 2640-3498.

Li, Z., Han, J., E, W., and Li, Q. (2021a). On the Curse of Memory in Recurrent Neural Networks: Approximation and Optimization Analysis. In *International Conference on Learning Representations 2021*. URL https://openreview.net/forum?id=8Sqhl-nF50.

Li, Z., Han, J., E, W., and Li, Q. (2022). Approximation and Optimization Theory for Linear Continuous-Time Recurrent Neural Networks. *Journal of Machine Learning Research*, 23(42), 1–85. URL http://jmlr.org/papers/v23/21-0368.html.

Li, Z., Jiang, H., and Li, Q. (2021b). On the approximation properties of recurrent encoder-decoder architectures. In *International Conference on Learning Representations 2022*. URL https://openreview.net/forum?id=xDIvIqQ3DXD.

Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*. URL http://arxiv.org/abs/1609.03499. ArXiv: 1609.03499.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Å., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. QID: Q30249683.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Construction of a Lyapunov function for linear coupled impulsive systems

Ivan Atamas, **Sergey Dashkovskiy**, **Vitalii Slynko**

Institute of Mathematics, University of Würzburg Germany, (e-mail: ivan.atamas, sergey.dashkovskiy, vitalii.slynko@uni-wuerzburg.de)

Abstract: We propose an approach to construct a Lyapunov function for a linear coupled impulsive system consisting of two time-invariant subsystems. In contrast to various variants of small-gain stability conditions for coupled systems, the asymptotic stability property of independent subsystems is not assumed. To analyze the asymptotic stability of a coupled system, the direct Lyapunov method is used in combination with the discretization method. The periodic case and the case when the Floquet theory is not applicable at all are considered separately.

Keywords: Impulsive systems, time-varying systems, coupled systems, Lyapunov methods

1. INTRODUCTION AND PROBLEM STATEMENT

Impulsive differential equations can model mechanical systems subjected to shocks. Instantaneous changes in holonomic or nonholonomic constraints imposed on the system and changes in the parameters of the system in time lead to the need to study stability of impulsive systems with variable coefficients. It is important to obtain stability conditions that are robust with respect to small variations in the sequence of moments of impulse action. A possible approach is to construct a matrix-valued Lyapunov function Djordjevic (1983); Martynyuk (1985). In Martynyuk and Slynko (2003) for linear time-variant coupled systems with time-invariant subsystems a construction of a matrix-valued Lyapunov function is proposed. For linear impulsive systems with variable coefficients the problem of choosing the elements of the matrix-valued Lyapunov function has not been studied.

The use of the discretization method to construct approximate solutions of the Lyapunov matrix differential equations has led to significant advances in the theory of stability of linear hybrid systems with constant parameters, see Allerhand and Shaked (2010). Discretization method for a construction of a Lyapunov function allows to obtain estimates of dwell-times that guarantee the stability of a linear hybrid system or conditions of robust stability.

We propose to apply the discretization method to construct matrix-valued Lyapunov functions for linear impulsive systems with periodic coefficients. We assume that independent subsystems are time-invariant and for the dwell-times two possible cases are considered: they are constant or subject to two-sided estimates. The elements of the matrix-valued Lyapunov function are constructed in the bilinear forms with time-variable matrices. The proposed construction of Lyapunov functions admits a numerical implementation.

Let $\mathbb{R}^n$ be the n-dimensional Euclidian space with standard dot product, $\mathbb{R}^{n\times m}$ be linear space of $n\times m$ matrices. For a matrix $A\in\mathbb{R}^{n\times n}$, $\sigma(A)$ denotes its spectrum, $r_\sigma(A)$ denotes spectral radius of A and norm $\|A\|=\lambda_{\max}^{1/2}(A^{\mathrm{T}}A)$. If $\sigma(A)\subset\mathbb{R}$, then $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are its smallest and largest eigenvalues respectively. For any symmetric matrices P and Q the notation $P\succeq Q$ means that $P-Q$ is a positive semidefinite matrix and $P\succ Q$ means that $P-Q$ is a positive definite matrix. We will use the Cauchy–Bunyakovskii inequality $|x^{\mathrm{T}}y|\le\|x\|\|y\|$ for $x,y\in\mathbb{R}^n$.

Consider a coupled linear system of impulsive differential equations consisting of two subsystems

$$\begin{aligned}\dot{x}_1(t)&=A_{11}x_1(t)+A_{12}(t)x_2(t),\quad t\neq\tau_k\\ \dot{x}_2(t)&=A_{21}(t)x_1(t)+A_{22}x_2(t),\quad t\neq\tau_k,\\ x_1(t^+)&=B_{11}x_1(t)+B_{12}x_2(t),\quad t=\tau_k,\\ x_2(t^+)&=B_{21}x_1(t)+B_{22}x_2(t),\quad t=\tau_k\end{aligned}\tag{1}$$

where $x_i\in\mathbb{R}^{n_i}$, $i=1,2$, $A_{ij}:\mathbb{R}\to\mathbb{R}^{n_i\times n_j}$ are piecewise continuous maps, $i,j=1,2$, A_{ii} are constant matrices and $A_{ij}(t)$, $i\neq j$ are θ-periodic functions, i.e., $A_{ij}(t+\theta)=A_{ij}(t)$ for all $t\in\mathbb{R}$, $\{\tau_k\}_{k=0}^{\infty}$ is a sequence of moments of impulse action, such that $\theta=\tau_k-\tau_{k-1}$, $k\ge1$. $B_{ij}\in\mathbb{R}^{n_i\times n_j}$ are constant matrices. We denote $x=(x_1^{\mathrm{T}},x_2^{\mathrm{T}})^{\mathrm{T}}$ and $n=n_1+n_2$. We will study the asymptotic stability of (1) in the sense of the following

Definition 1. *The linear impulsive system* (1) *is called*

1) stable if $\forall\varepsilon>0$, $t_0\in\mathbb{R}$ *there exists* $\delta=\delta(\varepsilon,t_0)>0$ *such that* $\|x_0\|<\delta\ \Rightarrow\ \|x(t,t_0,x_0)\|<\varepsilon$ *for all* $t\ge t_0$*;*

2) uniformly stable if $\forall\varepsilon>0\ \exists\delta=\delta(\varepsilon)>0$ *s.t.* $\forall\, t_0\in\mathbb{R}$ *we have* $\|x_0\|<\delta\ \Rightarrow\ \|x(t,t_0,x_0)\|<\varepsilon$ *for all* $t\ge t_0$*;*

3) asymptotically stable (AS) if it is stable and for all (t_0,x_0) *it holds that* $\lim\limits_{t\to+\infty}\|x(t,t_0,x_0)\|=0$.

Here, $x(t,t_0,x_0)$ is the solution to (1) with the initial condition $x(t_0,t_0,x_0)=x_0$, $x_0=(x_{10}^{\mathrm{T}},x_{20}^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^n$.

The aim of this work is to construct a Lyapunov function for (1). For simplicity we assume that $t_0=0$. Let

$$\mathbb{V}(t,x)=(v_{ij}(t,\cdot,\cdot))_{i,j=1,2}$$

be a matrix-valued Lyapunov function (MFL) Djordjevic (1983); Martynyuk (1985). We choose the diagonal elements of this function in the quadratic form $v_{ii}(t,x_i)=x_i^{\mathrm{T}}P_{ii}(t)x_i$, where $P_{ii}:\mathbb{R}\to\mathbb{R}^{n_i\times n_i}$, $P_{ii}(t)\succ0$, $i=1,2$ are continuous on the left θ-periodic maps, and off-diagonal elements in the form $v_{ij}(t,x_i,x_j)=x_i^{\mathrm{T}}P_{ij}(t)x_j$, where $P_{ij}:\mathbb{R}\to\mathbb{R}^{n_i\times n_j}$ are continuous on the left θ-periodic maps, $P_{ij}(t)=P_{ji}^{\mathrm{T}}(t)$. It is enough to define these functions $P_{ij}(t)$, $i,j=1,2$ on the period $(0,\theta]$.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2. MATRIX-VALUED LYAPUNOV FUNCTIONS

Discretization parameters are the number of nodes $N \in \mathbb{N}$ and the discretization step length $h = \frac{\theta}{N}$. Let $P_0 = (P_{ij}^{(0)})_{i,j=1,2}$ be a positive definite symmetric block matrix, $P_{ij}^{(0)} \in \mathbb{R}^{n_i \times n_j}$, $P_{ij}^{(0)} = (P_{ji}^{(0)})^{\mathrm{T}}$. We define recursively

$$P_{ij}^{(m)}, \quad P_{ji}^{(m)} = (P_{ij}^{(m)})^{\mathrm{T}}, \quad i,j = 1,2$$

as follows

$$P_{ii}^{(m+1)} = e^{-A_{ii}^{\mathrm{T}} h}(P_{ii}^{(m)} - \int_{mh}^{(m+1)h} (P_{ij}^{(m)} A_{ji}(s) + A_{ji}^{\mathrm{T}}(s) P_{ji}^{(m)})\, ds) e^{-A_{ii}h}, \quad i \neq j, \tag{2}$$

$$P_{12}^{(m+1)} = e^{-A_{11}^{\mathrm{T}} h}(P_{12}^{(m)} - \int_{mh}^{(m+1)h} (P_{11}^{(m)} A_{12}(s) + A_{21}^{\mathrm{T}}(s) P_{22}^{(m)})\, ds) e^{-A_{22}h}. \tag{3}$$

Next, define matrices $P_{ij}(t)$, $i,j = 1,2$, $P_{ij}(t) = P_{ji}^{\mathrm{T}}(t)$ on the intervals $(mh, (m+1)h]$ by setting

$$P_{ii}(t) = e^{-A_{ii}^{\mathrm{T}}(t-mh)}(P_{ii}^{(m)} - \int_{mh}^{t} (P_{ij}^{(m)} A_{ji}(s) + A_{ji}^{\mathrm{T}}(s) P_{ji}^{(m)})\, ds) e^{-A_{ii}(t-mh)}, \quad i \neq j, \tag{4}$$

$$P_{12}(t) = e^{-A_{11}^{\mathrm{T}}(t-mh)}(P_{12}^{(m)} - \int_{mh}^{t} (P_{11}^{(m)} A_{12}(s) + A_{21}^{\mathrm{T}}(s) P_{22}^{(m)})\, ds) e^{-A_{22}(t-mh)}. \tag{5}$$

We define an MLF $\mathbb{V}(t, x_1, x_2) = (v_{ij}(t, ., .))_{i,j=1,2}$, by $v_{ii}(t, x_i) = x_i^{\mathrm{T}} P_{ii}(t) x_i$, $i = 1,2$, $v_{ij}(t, x_i, x_j) = x_i^{\mathrm{T}} P_{ij}(t) x_j$, $i \neq j$, $P_{ij}(t) = P_{ji}^{\mathrm{T}}(t)$, $i,j = 1,2$. Using $\mathbb{V}$ we construct the scalar Lyapunov function Djordjevic (1983)

$$v(t, x_1, x_2) = v_{11}(t, x_1) + 2v_{12}(t, x_1, x_2) + v_{22}(t, x_2). \tag{6}$$

Assumption 1. *Let $\gamma_{12}^{(m)}$, $\gamma_{21}^{(m)}$, $m = 0, 1, \ldots, N-1$ be positive constants such that the following inequalities hold*

$$\sup_{s \in (mh, (m+1)h]} \|A_{12}(s)\| \leq \gamma_{12}^{(m)}, \quad \sup_{s \in (mh, (m+1)h]} \|A_{21}(s)\| \leq \gamma_{21}^{(m)}$$

There exists real constants μ_i, δ_i, M_i, N_i, such that for matrices A_{ii} the following estimates Gil (1993) holds

$$\|e^{sA_{ii}}\| \leq M_i e^{s\mu_i}, \quad \|e^{-sA_{ii}}\| \leq N_i e^{s\delta_i}, \quad s \geq 0.$$

To verify the positive-definiteness of the proposed LF and to construct the impulsive scalar equation we use

Lemma 1. *Let $z_{1m} = e^{-A_{11}(t-mh)} x_1$, $z_{2m} = e^{-A_{22}(t-mh)} x_2$, $t \in (mh, (m+1)h]$. Then,*

$$\lambda_{\min}(\Pi_m)\|z_m\|^2 \leq v(t, x_1, x_2) \leq \lambda_{\max}(\Xi_m)\|z_m\|^2, \quad \text{for all} \quad t \in (mh, (m+1)h], \quad m = 0, \ldots, N-1, \tag{7}$$

where $z_m = (z_{1m}^{\mathrm{T}}, z_{2m}^{\mathrm{T}})^{\mathrm{T}}$, $\|z_m\|^2 = \|z_{1m}\|^2 + \|z_{2m}\|^2$, $\Xi_m = (\xi_{ij}^{(m)})_{i,j=1,2}$ is some block matrix and $\Pi_m = (\pi_{ij}^{(m)})_{i,j=1,2}$ is block matrix with the elements

$$\pi_{11}^{(m)} = P_{11}^{(m)} - h(2\gamma_{21}^{(m)}\|P_{12}^{(m)}\| + (\gamma_{12}^{(m)}\|P_{11}^{(m)}\| + \gamma_{21}^{(m)}\|P_{22}^{(m)}\|))I_{n_1},$$
$$\pi_{22}^{(m)} = P_{22}^{(m)} - h(2\gamma_{12}^{(m)}\|P_{12}^{(m)}\| + (\gamma_{12}^{(m)}\|P_{11}^{(m)}\| + \gamma_{21}^{(m)}\|P_{22}^{(m)}\|))I_{n_2},$$
$$\pi_{12}^{(m)} = P_{12}^{(m)}, \quad \pi_{21}^{(m)} = P_{21}^{(m)}$$

For $i = 1, 2$, $i \neq j$ we denote

$$\eta_{ii}^{(m)} := \sqrt{\|P_{ii}^{(m)}\|^2 + \|P_{ij}^{(m)}\|^2} + \|P_{ij}^{(m)}\|,$$
$$\eta_{ij}^{(m)} := \frac{1}{2}(\|P_{ii}^{(m)}\|\gamma_{ij}^{(m)} + \|P_{jj}^{(m)}\|\gamma_{ji}^{(m)} + \sqrt{(\|P_{ii}^{(m)}\|\gamma_{ij}^{(m)} + \|P_{jj}^{(m)}\|\gamma_{ji}^{(m)})^2 + 16(\gamma_{ji}^{(m)})^2\|P_{ij}^{(m)}\|^2}).$$
$$\alpha_{11}^{(m)} := \gamma_{12}^{(m)}\|A_{22}\|N_1 M_2 \eta_{11}^{(m)}, \alpha_{12}^{(m)} := \gamma_{12}^{(m)}\|A_{11}\|N_1\eta_{11}^{(m)},$$
$$\alpha_{21}^{(m)} := \gamma_{21}^{(m)}\|A_{11}\|N_2 M_1 \eta_{22}^{(m)}, \alpha_{22}^{(m)} := \gamma_{21}^{(m)}\|A_{22}\|N_2\eta_{22}^{(m)},$$

$$\Theta_m(h) := \frac{\alpha_{11}^{(m)}}{\mu_2}\Big(\frac{e^{(\mu_2+\delta_1)h} - 1}{\mu_2 + \delta_1} - \frac{e^{\delta_1 h} - 1}{\delta_1}\Big) + \frac{\alpha_{12}^{(m)}}{\delta_1}\Big(\frac{e^{h\delta_1} - 1}{\delta_1} - h\Big) + \frac{\alpha_{21}^{(m)}}{\mu_1}\Big(\frac{e^{(\mu_1+\delta_2)h} - 1}{\mu_1 + \delta_2} - \frac{e^{\delta_2 h} - 1}{\delta_2}\Big) + \frac{\alpha_{22}^{(m)}}{\delta_2}\Big(\frac{e^{h\delta_2} - 1}{\delta_2} - h\Big) + \tag{8}$$
$$+2\Big(\gamma_{12}^{(m)}\eta_{12}^{(m)} N_1 M_2\Big(\frac{he^{(\mu_2+\delta_1)h}}{\mu_2 + \delta_1} - \frac{e^{(\mu_2+\delta_1)h} - 1}{(\mu_2 + \delta_1)^2}\Big) + \gamma_{21}^{(m)}\eta_{21}^{(m)} N_2 M_1\Big(\frac{he^{(\mu_1+\delta_2)h}}{\mu_1 + \delta_2} - \frac{e^{(\mu_1+\delta_2)h} - 1}{(\mu_1 + \delta_2)^2}\Big)\Big).$$

Lemma 2. *Suppose that for all $m = 0, \ldots, N-1$, the matrices Π_m are positive definite. Then*

$$v((m+1)h, x_1((m+1)h), x_2((m+1)h)) \leq e^{\frac{\Theta_m(h)}{\lambda_{\min}(\Pi_m)}} v(mh + 0, x_1(mh + 0), x_2(mh + 0)). \tag{9}$$

We establish sufficient conditions for (1) to be AS using the above $\mathbb{V}(t, x_1, x_2)$ and Lemma 2.

Theorem 1. *Let N be a positive natural number and $P_0 > 0$ be such that for $m = 0, \ldots, N-1$ the matrices Π_m are positive definite and it holds that*

$$Q := \sum_{m=0}^{N-1} \frac{\Theta_m(h)}{\lambda_{\min}(\Pi_m)} + \ln \lambda_{\max}(P_N^{-1} B^{\mathrm{T}} P_0 B) < 0.$$

Then the system (1) is asymptotically stable.

3. APPLICATION AND COMPARISON OF RESULTS

Here we compare Theorem 1 with known stability conditions for coupled time-varying systems. We restrict ourselves to a special case of high-frequency periodic functions that is θ is sufficiently small. The value of this parameter we obtain from Theorem 1, applied to the case $N = 1$. We will compare our results with small-gain conditions obtained on the basis of the ISS theory and the Lyapunov vector function. Moreover, we will also consider the case when one of the independent subsystems is not stable, and the known approaches from the theory of stability of coupled systems are not applicable.

Consider (1) and denote $\widehat{A}_{ij} = \frac{1}{\theta}\int_0^\theta A_{ij}(t)\,dt$ for $i \neq j$. Let $B = (B_{ij})_{i,j=1,2}$, $\widehat{A} = \begin{pmatrix} 0 & \widehat{A}_{12} \\ \widehat{A}_{21} & 0 \end{pmatrix}$ be block matrices. Consider the system of linear matrix inequalities

$$\begin{aligned} &\operatorname{diag}\{e^{\theta A_{11}^{\mathrm{T}}}, e^{\theta A_{22}^{\mathrm{T}}}\} B^{\mathrm{T}} P_0 B \operatorname{diag}\{e^{\theta A_{11}}, e^{\theta A_{22}}\} \\ &\qquad \prec P_0 - \theta(\widehat{A}^{\mathrm{T}} P_0 + P_0 \widehat{A}). \end{aligned} \tag{10}$$

Suppose it has a solution $P_0 = (P_{ij}^{(0)})_{i,j=1,2}$ in the form of a symmetric positive-definite matrix, i.e. $P_{ij}^{(0)} = (P_{ji}^{(0)})^{\mathrm{T}}$.

Let us define matrices

$$\pi_{11}^{(0)} = P_{11}^{(0)} - \theta(2\gamma_{21}^{(0)}\|P_{12}^{(0)}\| + (\gamma_{12}^{(0)}\|P_{11}^{(0)}\| + \gamma_{21}^{(0)}\|P_{22}^{(0)}\|))I_{n_1},$$
$$\pi_{22}^{(0)} = P_{22}^{(0)} - \theta(2\gamma_{12}^{(0)}\|P_{12}^{(0)}\| + (\gamma_{12}^{(0)}\|P_{11}^{(0)}\| + \gamma_{21}^{(0)}\|P_{22}^{(0)}\|))I_{n_2},$$
$$\pi_{12}^{(0)} = P_{12}^{(0)}, \quad \pi_{21}^{(0)} = P_{21}^{(0)}$$

and block matrix $P_1 = (P_{ij}^{(1)})_{i,j=1,2}$, $(P_{ij}^{(1)})^{\mathrm{T}} = P_{ji}^{(1)}$ with the blocks

$$P_{11}^{(1)} = e^{-A_{11}^{\mathrm{T}}\theta}(P_{11}^{(0)} - \theta(P_{12}^{(0)}\widehat{A}_{21} + \widehat{A}_{21}^{\mathrm{T}}P_{21}^{(0)}))e^{-A_{11}\theta}, \tag{11}$$

$$P_{22}^{(1)} = e^{-A_{22}^{\mathrm{T}}\theta}(P_{22}^{(0)} - \theta(\widehat{A}_{12}^{\mathrm{T}}P_{12}^{(0)} + P_{21}^{(0)}\widehat{A}_{12}))e^{-A_{22}\theta} \tag{12}$$

$$P_{12}^{(1)} = e^{-A_{11}^{\mathrm{T}}\theta}(P_{12}^{(0)} - \theta(P_{11}^{(0)}\widehat{A}_{12} + \widehat{A}_{21}^{\mathrm{T}}P_{22}^{(0)})))e^{-A_{22}\theta}. \tag{13}$$

Corollary 1. *If* $Q{:=}\frac{\Theta_0(\theta)}{\lambda_{\min}(\Pi_0)} + \ln\lambda_{\max}(P_1^{-1}B^{\mathrm{T}}P_0B) < 0$, $\Pi_0 = (\pi_{ij}^{(0)})_{i,j=1,2} > 0$, *then* (1) *is asymptotically stable.*

Example 1.

$$\begin{aligned} \dot{x}_1(t) &= 0.01x_1(t) - 0.2\sin^2\frac{2\pi t}{\theta}x_2(t), \\ \dot{x}_2(t) &= -0.1x_2(t) + 0.2\cos^2\frac{2\pi t}{\theta}x_1(t) \end{aligned} \tag{14}$$

Choose $\theta = 0.09$, $P_{11}^{(0)} = 18$, $P_{12}^{(0)} = P_{21}^{(0)} = -7$, $P_{22}^{(0)} = 12$. *It is obvious that* $\mu_1 = 0.01$, $\mu_2 = -0.1$, $\delta_1 = -0.01$, $\delta_2 = 0.1$, $\gamma_{12}^{(0)} = \gamma_{21}^{(0)} = 0.2$. *By direct calculations we get*

$$\Pi_0 = \begin{pmatrix} 17.208 & -7 \\ -7 & 11.208 \end{pmatrix},$$
$$P_1 = \begin{pmatrix} 18.09340255 & -7.002491081 \\ -7.002491081 & 12.08966719 \end{pmatrix}$$

and $\eta_{11}^{(0)} = 26.31320792$, $\eta_{22}^{(0)} = 20.89244399$, $\eta_{12}^{(0)} = 7.103656905$, $\eta_{21}^{(0)} = 7.103656905$, $\alpha_{11}^{(0)} = 0.5262641584$, $\alpha_{12}^{(0)} = 0.052626641584$, $\alpha_{21}^{(0)} = 0.04178488798$, $\alpha_{22}^{(0)} = 0.4178488798$, $Q = -0.00004279 < 0$. *Therefore, the system is asymptotically stable.*

At the same time, the first independent system is unstable, which makes it impossible to apply the small-gain theorem.

3.1 Comparison with the small-gain conditions

Consider (1), assuming $\int_0^\theta A_{ij}(t)\,dt = 0$ for $i \neq j$. Since the Lyapunov vector functions or small-gain results are only applicable when the independent subsystems are AS, we assume $r_\sigma(e^{\theta A_{ii}}B_{ii}) < 1$, $i = 1, 2$. Then (10) reduces to

$$e^{\theta A_{ii}^{\mathrm{T}}} B_{ii}^{\mathrm{T}} P_{ii} B_{ii} e^{\theta A_{ii}} \prec P_{ii}, \quad i = 1, 2. \tag{15}$$

To apply Theorem 1, we assume that $P_{ii}^{(0)} = P_{ii}$, $i = 1, 2$ $P_{12}^{(0)} = 0$, then from (11)–(13) we obtain $P_{12}^{(1)} = 0$ and

$$P_{11}^{(1)} = e^{-\theta A_{11}^{\mathrm{T}}} P_{11} e^{-\theta A_{11}}, \quad P_{22}^{(1)} = e^{-\theta A_{22}^{\mathrm{T}}} P_{22} e^{-\theta A_{22}}.$$

Let us define the matrix

$$\begin{aligned} \Phi = &\begin{pmatrix} e^{\theta A_{11}}P_{11}^{-1}e^{\theta A_{11}^{\mathrm{T}}} & 0 \\ 0 & e^{\theta A_{22}}P_{22}^{-1}e^{\theta A_{22}^{\mathrm{T}}} \end{pmatrix} \begin{pmatrix} B_{11}^{\mathrm{T}} & B_{21}^{\mathrm{T}} \\ B_{12}^{\mathrm{T}} & B_{22}^{\mathrm{T}} \end{pmatrix} \\ &\times \begin{pmatrix} P_{11} & 0 \\ 0 & P_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \end{aligned}$$

A consequence of Theorem 1 is the following Proposition.

Proposition 1. *Let us assume that for the system* (1) $\int_0^\theta A_{ij}(t)\,dt = 0$ *for* $i \neq j$ *and* $r_\sigma(\Phi) < 1$, *the conditions of Assumption 1 and the following inequalities hold*

$$\begin{aligned} &\theta < \min\Big\{\frac{\lambda_{\min}(P_{11})}{\varrho}, \frac{\lambda_{\min}(P_{22})}{\varrho}\Big\}, \\ &\frac{\Theta_0(\theta)}{\min\{\lambda_{\min}(P_{11}) - \theta\varrho, \lambda_{\min}(P_{22}) - \theta\varrho\}} \\ &\qquad < -\ln\lambda_{\max}(\Phi), \end{aligned} \tag{16}$$

where $\varrho = \gamma_{12}\|P_{11}\| + \gamma_{21}\|P_{22}\|$, *then* (1) *is AS.*

To compare the obtained results with known ones obtained on the basis of the ISS approach or Lyapunov vector function (small-gain conditions), we consider (1) without impulsive action, i.e. $B_{ii} = I$, $i = 1, 2$, $B_{ij} = 0$ for $j = 1, 2$, $i \neq j$ under the same assumption $\int_0^\theta A_{ij}(t)\,dt = 0$ for $i \neq j$.

Since the Lyapunov vector function or small-gain results are only applicable when the independent subsystems are asymptotically stable, we assume $\max\{\operatorname{Re}\lambda \,|\, \lambda \in \sigma(A_{ii})\} < 0$ for $i = 1, 2$. For given $Q_i \succ 0$, $i = 1, 2$ consider the linear algebraic Lyapunov equations

$$A_{ii}^{\mathrm{T}}P_{ii} + P_{ii}A_{ii} = -Q_i. \tag{17}$$

It is known that under our assumptions for the matrices A_{ii}, $i = 1, 2$ these equations have unique solutions in the form of symmetric positive-definite matrices P_{ii}.

Remark 1. *Solutions of matrix algebraic equations* (17) *satisfy linear matrix inequalities* (15), *up to* $O(\theta^2)$.

In this case matrix Φ is the following $\Phi = \operatorname{diag}\{e^{\theta A_{11}}P_{11}^{-1}e^{\theta A_{11}^{\mathrm{T}}}P_{11}, e^{\theta A_{22}}P_{22}^{-1}e^{\theta A_{22}^{\mathrm{T}}}P_{22}\}$. Since, θ is a small parameter, we can write

$$\Phi = I - \theta\operatorname{diag}\{P_{11}^{-1}Q_1, P_{22}^{-1}Q_2\} + O(\theta^2) \quad \Rightarrow$$

$$-\ln\lambda_{\max}(\Phi) = \theta\min\{\lambda_{\min}(P_{11}^{-1}Q_1), \lambda_{\min}(P_{22}^{-1}Q_2)\} + O(\theta^2)$$

is positive-defined for a sufficiently small $\theta > 0$. On the other hand, it is easy to show that $\Theta_0(\theta) = \frac{\Theta_0(\theta)}{\min\{\lambda_{\min}(P_{11}) - \theta\varrho, \lambda_{\min}(P_{22}) - \theta\varrho\}}$, hence $\exists\,\theta^* > 0$, such that for all $\theta \in (0, \theta^*)$ the conditions of Proposition 1 are satisfied.

Corollary 2. *For the system* (1) *without impulsive action with* $\int_0^\theta A_{ij}(t)\,dt = 0$ *exists* $\theta^* > 0$, *such that for all* $\theta \in (0, \theta^*)$ *the coupled system* (1) *is asymptotically stable.*

Note that θ^* is determined from the conditions (16).

We apply the same LF $V_1(x_1) = x_1^{\mathrm{T}}P_{11}x_1$, $V_2(x_2) = x_2^{\mathrm{T}}P_{22}x_2$ to study stability of (1) without impulsive action, using the small-gain theorem in Edwards et al. (2000). By Assumption 1 and the Cauchy-Bunyakovsky inequality we get the estimates of $\dot{V}_i$ along solutions of (1)

$$\begin{aligned} \dot{V}_i(x_i) &= -x_i^{\mathrm{T}}Q_ix_i + 2x_i^{\mathrm{T}}P_{ii}A_{ij}(t)x_j \\ &\le -\lambda_{\min}(P_{ii}^{-1}Q_i)V_i(x_i) \\ &+2\|P_{ii}\|^{1/2}\|P_{jj}\|^{-1/2}\gamma_{ij}V_i^{1/2}(x_i)V_j^{1/2}(x_j). \end{aligned} \tag{18}$$

Here $i \neq j$, $i, j = 1, 2$. To check the small-gain conditions from Edwards et al. (2000) (Theorem 4), we choose

$$\chi_i(r) = \Big(\frac{2\gamma_{ij}\|P_{ii}\|^{1/2}\|P_{jj}\|^{-1/2}}{\lambda_{\min}(P_{ii}^{-1}Q_i)} + \varepsilon\Big)^2 r.$$

This leads to the following sufficient condition for the asymptotic stability of (1) (small-gain condition)

$$\gamma_{12}\gamma_{21} < \frac{1}{4}\lambda_{\min}(P_{11}^{-1}Q_1)\lambda_{\min}(P_{22}^{-1}Q_2) \tag{19}$$

Remark 2. *The method of Lyapunov vector functions lead us to the same stability conditions. Indeed,* (18) *in the new variables* $y_i(t) = V_i^{1/2}(x_i(t))$ *leads to a linear system of differential inequalities*

$$\dot{y}_i(t) \le -\frac{1}{2}\lambda_{\min}(P_{ii}^{-1}Q_i)y_i(t) + \|P_{ii}\|^{1/2}\|P_{jj}\|^{-1/2}\gamma_{ij}y_j(t).$$

Application of the comparison principle leads to (19).

From (19) it follows that the small-gain conditions do not depend on θ. Therefore, it is possible to choose the parameters of the system (1), such that the conditions (19) are not satisfied, however, based on Corollary 1, this system is asymptotically stable for sufficiently small θ. We conclude that our approach leads to less conservative stability conditions than the known small-gain conditions.

Example 2. *Consider a second-order linear system*

$$\dot{x}_1(t) = -0.2x_1(t) + 0.15a_{12}(t)x_2(t),$$
$$\dot{x}_2(t) = -0.1x_2(t) + 0.2a_{21}(t)x_1(t)$$

where $a_{ij} \in C(\mathbb{R})$, $\|a_{ij}\|_{C[0,\theta]} = 1$, $\int_0^\theta a_{ij}(t)\,dt = 0$. *Here the small-gain condition* $0.03\|a_{12}\|_{C[0,\theta]}\|a_{21}\|_{C[0,\theta]} < 0.02$ *is not satisfied. We hvae* $\mu_1 = -0.2$, $\mu_2 = -0.1$, $\delta_1 = 0.2$, $\delta_2 = 0.1$.

Choose $\theta = 0.5$, $P_{11}^{(0)} = 2.5$, $P_{22}^{(0)} = 5$, $P_{12}^{(0)} = 0$, $M_1 = M_2 = N_1 = N_2 = 1$. *By direct calculation* $\gamma_{12}^{(0)} = 0.15$, $\gamma_{21}^{(0)} = 0.2$,

$$\Pi_0 = \begin{pmatrix} 1.8125 & 0 \\ 0 & 4.3125 \end{pmatrix}, P_1 = \begin{pmatrix} 3.053506895 & 0 \\ 0 & 5.525854586 \end{pmatrix}$$

Since $Q = -0.0050178428 < 0$, *the asymptotic stability of the considered system follows.*

4. EXAMPLES

Consider a linear fourth-order impulsive system (1) with

$$A_{11} = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}, \quad A_{22} = \begin{pmatrix} -1 & 0.01 \\ 0.01 & -1 \end{pmatrix},$$

$$A_{12}(t) = -2\sin^2(\omega t)I, \quad A_{21}(t) = 2\sin^2(\omega t)I,$$

$$B_{11} = \begin{pmatrix} 0.98 & 0 \\ 0 & 0.98 \end{pmatrix}, \quad B_{22} = \begin{pmatrix} 1.02 & 0 \\ -0.01 & 1.01 \end{pmatrix}$$
$$B_{12} = \begin{pmatrix} -0.01 & 0 \\ 0.02 & 0 \end{pmatrix}, \quad B_{21} = \begin{pmatrix} 0 & 0.01 \\ -0.05 & 0 \end{pmatrix}$$

Here, $\tau_{k+1} - \tau_k = \theta = 0.2$, $\omega = \frac{2\pi}{\theta}$. To check the asymptotic stability conditions obtained in Theorem 1, we choose $N = 50$, $P_{11}^{(0)} = 18I$, $P_{12}^{(0)} = -7I$, $P_{22}^{(0)} = 12I$. Then, $\min_{p=\overline{0,49}} \lambda_{\min}(\Pi_p) = 7.032226894$, $Q = -0.0464239887 < 0$. Therefore, the linear impulsive system (1) is asymptotically stable. Note that the independent subsystem

$$\dot{x}_2(t) = A_{22}x_2(t), \quad t \ne \tau_k,$$
$$x_2(t^+) = B_{22}x_2(t), \quad t = \tau_k, \tag{20}$$

is not stable due to the fact that $r_\sigma(e^{A_{22}\theta}B_{22}) > 1$.

Consider a linear impulsive system (1) with the matrices

$$A_{11} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A_{22} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$$

$$A_{12}(t) = \begin{pmatrix} 0.2\cos(\omega t) & -0.2\sin(\omega t) \\ 0.2\sin(\omega t) & 0.2\cos(\omega t) \end{pmatrix},$$

$$A_{21}(t) = \begin{pmatrix} 0.1\cos(\omega t) & -0.1\sin(\omega t) \\ 0.1\sin(\omega t) & 0.1\cos(\omega t) \end{pmatrix}$$

$$B_{11} = \begin{pmatrix} 1.2 & 0.1 \\ -0.1 & 1.5 \end{pmatrix}, \quad B_{22} = \begin{pmatrix} 0.5 & 0.05 \\ -0.05 & -0.5 \end{pmatrix}$$
$$B_{12} = \begin{pmatrix} 0.04 & 0.1 \\ 0.1 & 0.04 \end{pmatrix}, \quad B_{21} = \begin{pmatrix} 0.05 & 0.1 \\ 0.2 & 0.05 \end{pmatrix}$$

Here, $\tau_{k+1} - \tau_k = \theta = 0.5$, $\omega = \frac{2\pi}{\theta}$. To check the asymptotic stability conditions obtained in Theorem 1, we choose $N = 3$, $P_{11}^{(0)} = I$, $P_{12}^{(0)} = 0$, $P_{22}^{(0)} = I$. Then,

$$P_3 = \begin{pmatrix} 2.7171 & 0 & -0.00098 & -0.0189 \\ 0 & 2.7171 & 0.0189 & -0.00098 \\ -0.00098 & 0.0189 & 0.9024 & 0 \\ -0.0189 & -0.00098 & 0 & 0.9024 \end{pmatrix},$$

$\min_{p=\overline{0,2}} \lambda_{\min}(\Pi_p) = 0.851$, $Q = -0.11333 < 0$. Therefore (1) is asymptotically stable. Consider separately the continuous dynamics of a linear impulsive system, which is described by a linear time-variant system of ODEs

$$\dot{x}_1(t) = A_{11}x_1(t) + A_{12}(t)x_2(t),$$
$$\dot{x}_2(t) = A_{21}(t)x_1(t) + A_{22}x_2(t), \tag{21}$$

We denote

$$U(\omega t) = \begin{pmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{pmatrix}$$

and rewrite (21) as

$$\dot{x}_1(t) = -x_1(t) + 0.2U(\omega t)x_2(t),$$
$$\dot{x}_2(t) = 0.1U(\omega t)x_1(t) + 0.1x_2(t), \tag{22}$$

Consider the Lyapunov–Chetaev function $v(x_1, x_2) = 2\|x_2\|^2 - \|x_1\|^2$, the total derivative of which is

$$\dot{v}(x_1, x_2) = 2(0.2\|x_2\|^2 + \|x_1\|^2 + 0.4\sin(\omega t)x_1^T J x_2)$$
$$\ge 2(0.2\|x_2\|^2 + \|x_1\|^2 - 0.4\|x_1\|\|x_2\|) > 0,$$

for all $(x_1, x_2) \ne (0, 0)$, where $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

Therefore, (22) is unstable. Its discrete dynamics is also unstable since $r_\sigma(B) = 1.472162$.

REFERENCES

Allerhand, L.I. and Shaked, U. (2010). Robust stability and stabilization of linear switched systems with dwell time. *IEEE Transactions on Automatic Control*, 56(2), 381–386.

Djordjevic, M. (1983). Stability analysis of interconnected systems with possibly unstable subsystems. *Systems Control Lett.*, 3(3), 165–169.

Edwards, H.A., Lin, Y., and Wang, Y. (2000). On input-to-state stability for time varying nonlinear systems. In *Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No. 00CH37187)*, volume 4, 3501–3506. IEEE.

Gil, M. (1993). Estimate for the norm of matrix-valued functions. *Linear and Multilinear Algebra*, 35(1), 65–73.

Martynyuk, A.A. (1985). On the Lyapunov matrix-function and stability of motions. *Dokl. Akad. Nauk SSSR*, 280(5), 1062–1066.

Martynyuk, A.A. and Slynko, V.I. (2003). On the stability of motion of a large-scale system. *Differential Equations*, 39(6), 791–796.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Kron Reduction of Open Chemical Reaction Network: zero-moment matching and a priori error bound via generalized Gramian

M.A. Prawira Negara [*] **A.M. Burohman** [*,**]
B. Jayawardhana [*]

[*] *Engineering and Technology Institute Groningen, Faculty of Science and Engineering, University of Groningen, The Netherlands (e-mail: m.a.prawira.negara@rug.nl; a.m.burohman@rug.nl; b.jayawardhana@rug.nl).*
[**] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands*

Keywords: Kron-based model reduction, moment at zero, single-species single substrate chemical network, Gramian-based approach.

1. KRON REDUCTION METHOD FOR OPEN CRN WITH MASS-ACTION KINETICS

In this extended abstract, we propose a Kron-based model reduction method for open chemical reaction networks (CRN) with constant inflow and proportional outflow, which guarantees the preservation of network structures and interlacing property of the reduced-order model. Kron reduction was originally introduced for reduction of electrical circuit, see Dofler and Bullo (2012). The concept of Kron reduction is similar with singular perturbation model reduction presented in Liu and Anderson (1989) in which the Schur complement is used.

Consider an open CRN with the dynamic given by

$$\left.\begin{aligned} \dot{x} &= ZDv(x) + ZD_{\text{in}}v_{\text{in}} - ZD_{\text{out}}v_{\text{out}}(x) \\ y &= C\text{Exp}(Z^T\text{Ln}x), \end{aligned}\right\} \tag{1}$$

where $Z \in \mathbb{R}_+^{n\times c}$ is the complex stoichiometric matrix of the network, $D \in \mathbb{R}^{c\times r}$ is the incidence matrix and $v(x) \in \mathbb{R}^r$ is the vector of reaction rates or fluxes, D_{in} and D_{out} are incidence matrices of the inflow and outflow that connect internal complexes to an additional "zero"-complex $\emptyset$, $v_{\text{in}} \in \mathbb{R}^c$ is the vector of inflow from the environment, $v_{\text{out}}(x) \in \mathbb{R}^d$ is gives the outflow kinetics and y is the measured output. We will assume throughout that the inflow v_{in} are constant inflow while the outflow kinetics $v_{\text{out}}(x)$ and the vector of reaction rates $v(x) \in \mathbb{R}^r$ are given by mass-action kinetics, as presented in van der Schaft et al. (2016), the reaction rates are given by

$$v(x) = K\text{Exp}(Z^T\text{Ln}x), \tag{2}$$

$$v_{\text{out}}(x) = K_{\text{out}}\text{Exp}(Z^T\text{Ln}x), \tag{3}$$

where the *outgoing co-incidence matrix* $K \in \mathbb{R}^{r\times c}$ is the matrix whose (j,σ)th element equals the j-th reaction rate constant $k_j > 0$ if the σ-th complex is the substrate complex for the j-th reaction and K_{out} is the *outgoing co-incidence matrix* for the outflow. This allow us to rewrite (1) into the form of

$$\left.\begin{aligned} \dot{x} &= -Z(L+R)\text{Exp}(Z^T\text{Ln}x) + ZD_{\text{in}}v_{\text{in}} \\ y &= C\text{Exp}(Z^T\text{Ln}x), \end{aligned}\right\} \tag{4}$$

where $L := -DK \in \mathbb{R}^{c\times c}$ defines a weighted Laplacian matrix which has non-negative diagonal elements and non-positive off-diagonal elements and $R = D_{\text{out}}K_{\text{out}}$. Let us use the partitions

$$Z = [Z_1 \ Z_2], \quad L = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix},$$
$$R = \begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix}, \text{ and } D_{\text{in}} = \begin{bmatrix} D_{\text{in},1} \\ D_{\text{in},2} \end{bmatrix}. \tag{5}$$

For ease of expression, we consider the following auxiliary dynamical system

$$\begin{bmatrix} \dot{\xi}_1 \\ \dot{\xi}_2 \end{bmatrix} = -\begin{bmatrix} L_{11}+R_{11} & L_{12} \\ L_{21} & L_{22}+R_{22} \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} D_{\text{in},1}v_{\text{in}} \\ D_{\text{in},2}v_{\text{in}} \end{bmatrix}, \tag{6}$$

which corresponds to the dynamics of complexes in (4) with $w_1 = \text{Exp}(Z_1^T\text{Ln}x)$ and $w_2 = \text{Exp}(Z_2^T\text{Ln}x)$. By imposing the constraint $\dot{\xi}_2 = 0$, we obtain that the reduced network is given by

$$\left.\begin{aligned} \dot{x} &= -\hat{Z}\hat{L}\text{Exp}(\hat{Z}^T\text{Ln}x) + \hat{Z}\hat{D}_{\text{in}}v_{\text{in}} \\ y &= \hat{C}\text{Exp}(\hat{Z}^T\text{Ln}x), \end{aligned}\right\} \tag{7}$$

where $\hat{Z} = Z_1$, $\hat{L} = (L_{11}+R_{11}) - L_{12}(L_{22}+R_{22})^{-1}L_{21}$, $\hat{D}_{\text{in}} = D_{\text{in},1} - L_{12}(L_{22}+R_{22})^{-1}D_{\text{in},2}$, and $\hat{C} = C_1 - C_2(L_{22}+R_{22})^{-1}L_{21}$.

2. ANALYSIS OF KRON-REDUCED OPEN CRN

To gain further insight to the dynamics of Kron-reduced open CRN, we will investigate a number of dynamical properties that can be preserved or obtained by the resulting Kron-reduced open CRN. First, we will see the interlacing property, where, in van der Schaft et al. (2013), Rao et al. (2013) and Jayawardhana et al. (2015), it has been shown that the Kron reduction approach preserve the network structure of the original CRN. For instance, if

the original CRN is detailed-balanced or complex-balanced then the Kron-reduced CRN is again detailed-balanced or complex-balanced, respectively. Another network property that is inherited by the Kron-reduced CRN is the network spectrum interlacing property where the spectrum of weighted Laplacian matrix of the Kron-reduced CRN is interlaced with that of the original CRN. It follows that for a given detailed-balanced open CRN as in (4), consider the corresponding Kron-reduced open CRN as in (7). Then $\sigma(\hat{L})$ interlace with $\sigma(L+R)$, i.e. for every $i = 1, \ldots, \hat{c}$

$$0 < \lambda_i(L+R) \leq \lambda_i(\hat{L}) \leq \lambda_{i+c-\hat{c}}(L+R), \tag{8}$$

holds. The proof of the proposition follows the standard result for Kron reduction of a positive semi-definite Hermitian matrix as in Smith (1992) that is used for electrical networks in Dofler and Bullo (2012) or closed CRN in Jayawardhana et al. (2015).

The next thing that we will investigate is the steady-state or zero-moment matching property. Suppose that $\mathrm{Ker}(Z) = \emptyset$ and the underlying CRN graph $\mathcal{G}$ is undirected and connected. Then the zero-moment of reduced open CRN in (7) matches with the zero-moment of original open CRN in (4). Since Z has full column rank, the left-inverse (or Moore-Penrose inverse) of Z is given by $Z^\dagger = (Z^T Z)^{-1} Z^T$. Pre-multiplying the first equation in (4) by $Z^\dagger$, we have the zero-moment property that satisfies

$$\left.\begin{aligned} 0 &= -A\mathrm{Exp}(Z^T \mathrm{Ln} x) + D_{\mathrm{in}} v_{\mathrm{in}} \\ y &= C\mathrm{Exp}(Z^T \mathrm{Ln} x), \end{aligned}\right\} \tag{9}$$

where $A = L + R$. Similarly, for the reduced-order open CRN in (7), its zero-moment satisfies

$$\left.\begin{aligned} 0 &= -(A_{11} - A_{12}A_{22}^{-1}A_{21})\mathrm{Exp}(Z^T \mathrm{Ln} x) \\ &\quad + \Big(D_{\mathrm{in},1} - (A_{12})(A_{22})^{-1} D_{\mathrm{in},2}\Big) v_{\mathrm{in}} \\ y &= (C_1 - C_2(A_{22})^{-1}(A_{21}))\mathrm{Exp}(\hat{Z}^T \mathrm{Ln} x). \end{aligned}\right\} \tag{10}$$

By hypotheses of the proposition, the matrix A is invertible due to the connectedness of $\mathcal{G}$ and due to the fact that R is a diagonal matrix with at least one positive entry (see, for example, Lemma 3 in Ni and Cheng (2010)). It follows that

$$y = CA^{-1} D_{\mathrm{in}} v_{\mathrm{in}}. \tag{11}$$

For the Kron-reduced one in (10), we can have a similar expression as above. In which we will have

$$y = \hat{C}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} \hat{D}_{\mathrm{in}} v_{\mathrm{in}}, \tag{12}$$

where $\hat{C}$ and $\hat{D}_{in}$ are as in (7). By using matrix inversion lemma or Woodbury formula (see Riedel (1992)), it follows that (11) is equivalent to (12).

3. SELECTION OF REMOVED COMPLEXES OF OPEN SS CRN VIA GENERALIZED GRAMIANS

As shown in Rao et al. (2014), the selection of removed nodes using Kron reduction method in a closed CRN plays an important role in the quality of the approximation error. Correspondingly, Rao et al. (2014) has proposed the combined use of error integral and simulation to remove one node at a time in order to obtain the set of removed nodes. Since the computing resource that is needed to perform the said method is quite extensive, we will propose another method with the use of generalized Gramian to get the optimal set of removed nodes along with the model reduction error bound for a class of open detailed-balanced single-species single-substrate (which we will refer to as SS) CRN. The open SS CRN can be given by

$$\left.\begin{aligned} \dot{x} &= \underbrace{-(\overbrace{DK}^{=:L} + \overbrace{D_{\mathrm{out}}K_{\mathrm{out}}}^{=:R})}_{=:A} x + \underbrace{D_{\mathrm{in}}}_{=:B} u \\ y &= Cx \end{aligned}\right\} \tag{13}$$

where the matrices A, B and C are the usual matrices of linear systems.

Since $Z = I$ in this case, the partitioning of Z, L, R and D_{in} as above corresponds to the partitioning of matrices A, B and C in (13) as follows

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = [C_1 \; C_2]. \tag{14}$$

Hence the application of Kron reduction method to (13) gives

$$\left.\begin{aligned} \dot{\hat{x}} &= \hat{A}\hat{x} + \hat{B}u \\ \hat{y} &= \hat{C}\hat{x}, \end{aligned}\right\} \tag{15}$$

where $\hat{A} = A_{11} - A_{12}A_{22}^{-1}A_{21}$, $\hat{B} = B_1 - A_{12}A_{22}^{-1}B_2$, and $\hat{C} = C_1 - C_2A_{22}^{-1}A_{21}$.

For linear systems, controllability and observability Gramians have been used to obtain reduced-order models, see, e.g. Antoulas (2005). These Gramians reveal the states of systems that are hard to control and observe. Instead of using the ordinary Gramian to get the controllability or observability Gramian, generalized Gramians can be defined to characterize state variables that are difficult to control or to observe. In particular, *generalized controllability Gramians* is defined as a solution of inequality

$$AP + PA^T + BB^T \leq 0, \tag{16}$$

and, similarly, *generalized observability Gramians* is a solution of inequality

$$A^T Q + QA + C^T C \leq 0. \tag{17}$$

Note that, the matrices P and Q in (16) and (17) are not unique and satisfy $P \geq P_0$ and $Q \geq Q_0$ with P_0 and Q_0 be the ordinary controllability and observability Gramian, respectively, which are the unique solutions of their corresponding Lyapunov equations. This non-uniqueness gives extra degree of freedom on their structure. Namely, we can force P and Q to have a specific structure, such as forcing P and Q to be diagonal.

In our method, we will have a slightly different definition of generalized Gramians. Namely, matrices $P \in \mathbb{R}_+^{n\times n}$ and $Q \in \mathbb{R}_+^{n\times n}$ are said to be *generalized controllability* and *observability Gramians of open SS CRN systems* (13) if they are diagonal and satisfy

$$AP + PA^T + BB^T \leq 0 \tag{18}$$

and

$$A^T Q + QA + A^T C^T CA \leq 0, \tag{19}$$

respectively, where A, B and C are as in (13).

We remark that the matrix inequality defined above is stronger than the one defined in (17). It can be verified that if Q^* is a solution of (17) then $A^T Q^* A$ is a solution of (19). The generalized Gramians of open SS CRN defined above will allow for the computation of error bounds below. In this regards, the computation of tight model reduction error bounds via (18) and (19) can be done by minimizing

trace(P) and trace(Q). Following this definition we can express the generalized Gramians as

$$P = \begin{bmatrix} \pi_1^c & & \\ & \ddots & \\ & & \pi_n^c \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} \pi_1^o & & \\ & \ddots & \\ & & \pi_n^o \end{bmatrix}. \tag{20}$$

Next, let us first consider a one step Kron reduction, where we only remove one complex that is deemed the least controllable and observable from generalized Gramian standpoint, as follows. Consider an open SS CRN system Σ as in (13) and its reduced-order model $\hat{\Sigma}_{n-1}$ via Kron reduction as in (15) by removing the n-th node so that the reduced system $\hat{\Sigma}_{n-1}$ are given by system matrices $\hat{A} = A_{11} - A_{12}A_{22}^{-1}A_{21}$, $\hat{B} = B_1 - A_{12}A_{22}B_2$ and $\hat{C} = C_1$. In this case, we assume that $C_2 = 0$. This corresponds to the assumption that we will not remove the nodes that are measured directly. Then

$$\hat{P}_1 = \begin{bmatrix} \pi_1^c & & \\ & \ddots & \\ & & \pi_{n-1}^c \end{bmatrix} \quad \text{and} \quad \hat{Q}_1 = \begin{bmatrix} \pi_1^o & & \\ & \ddots & \\ & & \pi_{n-1}^o \end{bmatrix}, \tag{21}$$

are generalized controllability and observability Gramians for system $\hat{\Sigma}_{n-1}$, respectively.

Moreover, we have also proven that the ignored entries of the Gramians can be used to compute an a priori upper bound of this one step model reduction error as follows. For any input function $u(\cdot) \in \mathcal{L}_2[0,\infty)$ and initial condition $x(0) = 0$ and $\hat{x}(0) = 0$, the outputs satisfy

$$\|y - \hat{y}\|_2 \leq 2M_{22}\sqrt{(\pi_n^c \pi_n^o)}\|u\|_2, \tag{22}$$

where the scalar $M_{22} > 0$ is a diagonal element of the partition matrix

$$M := \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = -A^{-1} = (L+R)^{-1}, \tag{23}$$

and π_n^c, π_n^o are the corresponding removed elements of the generalized Gramians P and Q as in (20). The proof of proposition is based on that of (Besselink et al., 2015, Theorem 11).

Based on the bound (22), we can order the complexes (or vertices of CRN) such that

$$M_{11}^2 \pi_1^c \pi_1^o \geq \cdots \geq M_{nn}^2 \pi_n^c \pi_n^o \geq 0. \tag{24}$$

Based on this order, we can consider the removal of complexes associated to smallest error bound. By removing the vertex corresponding to the smallest error bound, we can guarantee that the reduced-order model will have a small approximation error, but not necessarily the smallest. Note that such ordering procedure corresponds simply to applying a coordinate transformation Tx using a permutation matrix T.

In practice, when we apply our Kron reduction method to a CRN, we need to truncate not only one complex. From the a priori upper bound (22), we can extend this bound for the truncation of a set of complexes. By mainly utilizing triangle inequality, we have that for any input $u \in \mathcal{L}_2[0,\infty)$ and initial condition $x(0) = 0$ and $\hat{x}(0) = 0$, the outputs y and $\hat{y}_r$ satisfy the bound

$$\|y - \hat{y}_r\|_2 \leq 2\left(\sum_{i=r+1}^{n} M_{ii}\sqrt{\pi_i^c \pi_i^o}\right)\|u\|_2, \tag{25}$$

where y and $\hat{y}_r$ are the outputs of the original and the reduced-order model, respectively, π_i^cs and π_i^cs are the removed generalized controllability and observability Gramians, respectively, as in (20), and the scalar M_{ii} is the i-th diagonal element of the matrix $M = -A^{-1} = (L + R)^{-1}$.

4. CONCLUSIONS

We studied the usage of Kron Reduction in a balanced biochemical reaction network. Our study shows that the use of the Kron Reduction method is able to preserve the network structure and equilibrium point in the reduced network even with the presence of inflow and outflow in the network. Moreover, we also provide an a priori upper bound of the approximation error via generalized Gramian approach. The latter property has allowed us to guide systematically the selection of removed nodes/species via Kron reduction.

REFERENCES

Antoulas, A.C. (2005). *Approximation of Large-Scale Dynamical Systems.* Advances in Design and Control, SIAM, Philadelphia, PA.

Besselink, B., Sandberg, H., and Johansson, K.H. (2015). Clustering-based model reduction of networked passive systems. *IEEE Trans. on Automatic Control*, 61(10), 2958–2973.

Dofler, F. and Bullo, F. (2012). Kron reduction of graphs with applications to electrical networks. *IEEE Trans. on Circuits and Systems I: Regular Papers*, 60, 150–163.

Jayawardhana, B., Rao, S., Sikkema, W., and Bakker, B.M. (2015). Handling biological complexity using Kron reduction. *Mathematical Control Theory I*, 73–93.

Liu, Y. and Anderson, B.D.O. (1989). Singular perturbation approximation of balanced systems. *International Journal of Control*, 50(4), 1379–1405.

Ni, W. and Cheng, D. (2010). Leader-following consensus of multi-agent systems under fixed and switching topologies. *Systems & Control Letters*, 59, 209–217.

Rao, S., van der Schaft, A.J., and Jayawardhana, B. (2013). A graph-theoretical approach for the analysis and model reduction of complex-balanced chemical reaction networks. *Journal of Mathematical Chemistry*, 51, 2401–2422.

Rao, S., van der Schaft, A.J., van Eunen, K., Bakker, B.M., and Jayawardhana, B. (2014). A model reduction method for biochemical reaction networks. *BMC Systems Biology*, 8.

Riedel, K.S. (1992). A Sherman–Morrison–Woodbury identity for rank augmenting matrices with application to centering. *SIAM Journal on Matrix Analysis and Applications*, 13-2, 659–662.

Smith, R.L. (1992). Some interlacing properties of the Schur complement of a Hermitian matrix. *Linear Algebra and its Applications*, 177, 137–144.

van der Schaft, A.J., Rao, S., and Jayawardhana, B. (2013). On the mathematical structure of balanced chemical reaction networks governed by mass action kinetics. *Siam Journal on Applied Mathematics*, 953–973.

van der Schaft, A.J., Rao, S., and Jayawardhana, B. (2016). A network dynamics approach to chemical reaction networks. *International Journal of Control*, 89-4, 731–745.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Extended Abstract: Discovering collective variable dynamics of agent-based models

Marvin Lücke [*] **Péter Koltai** [**] **Stefanie Winkelmann** [*]
Nora Molkenthin [***] **Jobst Heitzig** [***]

[*] *Zuse Institute Berlin, Germany.*
[**] *Freie Universität Berlin, Germany.*
[***] *Potsdam Institute for Climate Impact Research, Germany.*

Abstract: Analytical approximations of the macroscopic behavior of agent-based models (e.g. via mean-field theory) often introduce a significant error, especially in the transient phase. For an example model called *continuous-time noisy voter model*, we use two data-driven approaches to learn the evolution of collective variables instead. The first approach utilizes the SINDy method to approximate the macroscopic dynamics without prior knowledge, but has proven itself to be not particularly robust. The second approach employs an informed learning strategy which includes knowledge about the agent-based model. Both approaches exhibit a considerably smaller error than the conventional analytical approximation.

Keywords: Opinion Dynamics (91D30), System Identification (93B30), Linear regression (62J07), Large Population Limit, Continuous-time Markov Chain.

1. INTRODUCTION

In the past years *agent-based modeling* has undoubtedly gained in importance. While its success in various applications is significant, agent-based models (ABMs) still have many limitations. Even for rather simple interaction rules between agents, the emerging macroscopic system behavior may be very complex and hard to predict analytically. Thus, a formal analysis is typically out of reach and computer simulations are used to examine the behavior of ABMs instead. As simulations of most ABMs scale at least linearly with the number of agents, they are often computationally infeasible for large populations of agents. Thus, we are interested in finding low-dimensional model reductions that scale sublinearly while preserving the macroscopic behavior of the underlying ABMs.

It is known that this macroscopic behavior can sometimes be expressed by a small number of *collective variables* (or *reaction coordinates*, to borrow a term from statistical physics and computational chemistry) that aggregate the most important dynamical information, see for example Helfmann et al. (2021). Furthermore, it is often observed that for a large number of agents the *effective dynamics* of the collective variables follows an almost deterministic and smooth evolution. Hence, the macroscopic dynamics may be approximated by a differential equation of the collective variables. We we will refer to this phenomenon as a *concentration (of measure) effect*. If we are only interested in the aggregated information provided by the collective variables, we can use the reduced system instead of the full agent-based model, which greatly lowers computational cost.

Finding suitable collective variables for a given system is a challenging task, see for instance Bittracher et al. (2017). However, we will assume from now on that a set of collective variables has been chosen and focus on the question of discovering their dynamics instead, i.e., finding a description of their evolution in the form of a differential equation.

In the next section we introduce an exemplary ABM called the *continuous-time noisy voter model* (CNVM), which models simple spreading processes on networks. The remaining sections of this extended abstract deal with analyzing this model for different examples of network topologies. In section 3 we present results about the CNVM on fully connected networks, where the appearance of concentration effects can be proven. In section 4 we discuss a ring-shaped network topology and derive approximate macroscopic dynamics analytically as well as via data-driven methods.

Although the network topologies that we discuss in sections 3 and 4 are rather simple, the methods we present lay the groundwork for analyzing more complex systems, which are more similar to real-world social networks and will be subject matter of future works.

2. THE CONTINUOUS-TIME NOISY VOTER MODEL

We define the continuous-time noisy voter model as a dynamical system on an undirected simple graph G with N nodes. Each node $i \in \{1, \dots, N\}$ in the graph represents an agent and has one of M opinions $\{1, \dots, M\}$. The state of the system at times $t \in \mathbb{R}_{\geq 0}$ is given by the stochastic process $\boldsymbol{x}(t) \in \mathcal{X}_N := \{1, \dots, M\}^N$, where $\boldsymbol{x}_i(t)$ is the opinion of node i at time t. We model the switching of the agents between opinions as Markov jump processes, i.e., $\boldsymbol{x}_i(t)$ is a continuous-time Markov chain on $\{1, \dots, M\}$ with transition rate matrix (generator matrix) $Q_i^G(\boldsymbol{x}) \in \mathbb{R}^{M,M}$. We define the rate at which agent i transitions from opinion m to a different opinion n as

$$(Q_i^G(\boldsymbol{x}))_{mn} = r_{mn}\frac{k_{i,n}^G(\boldsymbol{x})}{k_i^G} + \tilde{r}_{mn}, \tag{1}$$

where $r_{mn}, \tilde{r}_{mn} \in \mathbb{R}_{\geq 0}$ are constant parameters, k_i^G is the degree of the i-th node, and $k_{i,n}^G$ is the number of adjacent nodes (neighbors) of agent i that have opinion n. The first term of the rate (1) describes an influence on agent i by neighboring nodes of opinion n, while the second term controls transitions independently of the neighborhood of agent i, i.e., noise.

The CNVM as described above is a simple but quite general example of a spreading process on a network. The behavior of this model is heavily dependent on the underlying network G. Let us first examine the easiest case: the fully connected network.

3. FULLY CONNECTED NETWORK

Let us discuss the continuous-time noisy voter model in the case of a fully connected graph G_N on N nodes. Note that the cumulative rate of transitions from opinion m to opinion n happening is given by

$$\begin{aligned}\alpha_{mn}^{(N)}(\boldsymbol{x}) &:= \sum_{i:x_i=m} (Q_i^{G_N}(\boldsymbol{x}))_{mn} \\ &= \boldsymbol{y}_m\Big(r_{mn}\frac{\boldsymbol{y}_n}{N-1} + \tilde{r}_{mn}\Big),\end{aligned} \tag{2}$$

where the stochastic process $\boldsymbol{y}(t) \in \mathbb{N}_0^M$ is the collection of *opinion counts*, i.e.,

$$\boldsymbol{y}_m := \#\{i \in \{1,\dots,N\} \mid \boldsymbol{x}_i = m\}. \tag{3}$$

Due to their appearance in the *propensity function* $\alpha_{mn}^{(N)}$ in (2), the opinion counts $\boldsymbol{y}$ present themselves as a natural choice of collective variables. In fact, we are able to completely describe the ABM dynamics using $\boldsymbol{y}$, see Winkelmann and Schütte (2020).

Moreover, it is known that the evolution of *opinion shares* $\boldsymbol{c} := \frac{1}{N}\boldsymbol{y}$ shows a concentration effect, which is described in the following theorem.

Theorem 1. (Kurtz (1978)) Let $(N_\ell)_{\ell\in\mathbb{N}} \subset \mathbb{N}$ be strictly increasing and let $\boldsymbol{c}^{(\ell)}(t)$ denote the stochastic process of opinion shares on the fully connected graph on N_ℓ nodes. Assume it holds $\boldsymbol{c}^{(\ell)}(0) = c_0$ almost surely (a.s.) for all ℓ. Then

$$\boldsymbol{c}^{(\ell)}(t) \xrightarrow[\ell\to\infty]{a.s.} c(t) \quad \text{for all } t \in \mathbb{R}_{\geq 0},$$

where c(t) is the solution of the *reaction-rate equation* (RRE)

$$\dot{c}(t) = \sum_{\substack{m,n=1\\ m\neq n}}^{M} \tilde{\alpha}_{mn}(c(t))(e_n - e_m)\ , \quad c(0) = c_0,$$

e_n, e_m are standard unit vectors and

$$\tilde{\alpha}_{mn}(c) := c_m\Big(r_{mn}c_n + \tilde{r}_{mn}\Big).$$

An illustration on how the stochastic processes $\boldsymbol{c}^{(\ell)}(t)$ become more concentrated around the solution of the RRE as the number of agents increases can be found in Figure 1. In the next section we will examine a more challenging network topology.

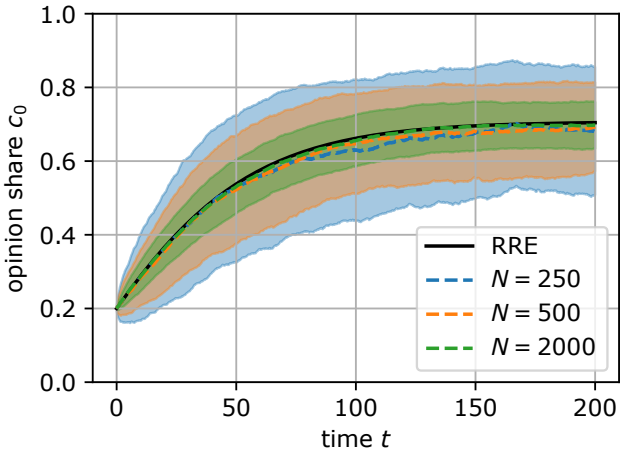


Fig. 1. Mean (dashed line) ± one standard deviation (shaded area) of 500 realizations of the CNVM on fully connected networks of sizes $N = 250, 500, 2000$. Black: solution of the reaction-rate equation.

4. RING-SHAPED NETWORK

While the ring-shaped network (cycle graph) still has a fairly simple structure, analytical results for concentration effects with respect to suitable collective variables are not available. We assume that there are only $M = 2$ opinions, 0 and 1, and set $m \in \{0, 1\}$ and $n = 1 - m$. Let us again inspect the propensity function for this network structure:

$$\begin{aligned}\alpha_{mn}^{(N)}(\boldsymbol{x}) &= \sum_{i:x_i=m} (Q_i^{G_N}(\boldsymbol{x}))_{mn} \\ &= \frac{r_{mn}}{2}(\boldsymbol{y}_m^1 + 2\boldsymbol{y}_m^2) + \boldsymbol{y}_m\tilde{r}_{mn},\end{aligned} \tag{4}$$

where $\boldsymbol{y}$ is defined as in (3), and

$$\boldsymbol{y}_m^\ell := \#\{i \in \{1,\dots,N\} \mid \boldsymbol{x}_i = m \text{ and } k_{i,n}^G(\boldsymbol{x}) = \ell\} \tag{5}$$

is the count of agents of opinion m with ℓ neighbors of opinion $n = 1 - m$. Thus, a natural choice of collective variables is given by

$$\boldsymbol{z} := (\boldsymbol{y}_0^0, \boldsymbol{y}_0^1, \boldsymbol{y}_0^2, \boldsymbol{y}_1^0, \boldsymbol{y}_1^1, \boldsymbol{y}_1^2). \tag{6}$$

Note that $\boldsymbol{z}$ includes $\boldsymbol{y}$ because $\boldsymbol{y}_m = \boldsymbol{y}_m^0 + \boldsymbol{y}_m^1 + \boldsymbol{y}_m^2$.

4.1 The analytical approach

In this section we will derive an ODE $\dot{z} = f(z)$ that approximates the evolution of the collective variables $\boldsymbol{z}$. For this purpose we need to characterize at which rate certain opinion transitions happen and how they impact $\boldsymbol{z}$.

Let us discuss the following example: $(x_{i-1}, x_i, x_{i+1}) = (m, m, m)$ and agent i transitions from opinion m to opinion n. This transition has a rate of $\tilde{r}_{mn}$ (cf. (1)), because agent i has no neighbor of different opinion, who could influence them. Due to the transition of agent i, y_m^0 decreases and y_n^2 increases by one. Additionally, for the agents $(i-1)$ and $(i+1)$ the number of neighbors with opinion n changes as well, but this depends on the opinions of agents $(i-2)$ and $(i+2)$. Assume we knew that $(x_{i-2}, x_{i+2}) = (m, m)$. Then the transition of agent i would further increase y_m^1 by two and decrease y_m^0 by two, leading to the total *state-change vector* $v = (-3, 2, 0, 0, 0, 1)$. The

three other cases $(x_{i-2}, x_{i+2}) = (m,n), (n,m), (n,n)$ yield three (potentially) different *reactions* with different state-change vectors. Unfortunately, only the combined rate $\tilde{r}_{mn}$ of all four reactions is known, not the single reaction rates. We distribute the total rate onto the four reactions by relative occurrence and assuming stochastic independence of x_{i-2} and x_{i+2}. For the above example $(x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}) = (m,m,m,m,m)$ this yields a share of

$$\Big(\frac{y^0_m}{y^0_m + 0.5y^1_m}\Big)^2$$

of the combined rate $\tilde{r}_{mn}$. Hence, we set the rate of the above reaction given by $v = (-3,2,0,0,0,1)$ happening somewhere in the system (propensity) to

$$\alpha(z) = y^0_m \tilde{r}_{mn} \Big(\frac{y^0_m}{y^0_m + 0.5y^1_m}\Big)^2. \tag{7}$$

Following this procedure for all possible scenarios, i.e., agent i being of opinion $m = 0,1$ and having $\ell = 0,1,2$ neighbors of opinion n, and agents $(i-2), (i+2)$ having opinions $(m,m), (m,n), (n,m), (n,n)$, yields a total of $2 \cdot 3 \cdot 4 = 24$ reactions with associated reaction rates $\alpha_j(z)$ and state-change vectors v_j, $j = 1, \dots, 24$. Thus, the evolution of z is given by

$$\dot{z} = \sum_{j=1,\dots,24} \alpha_j(z) v_j. \tag{8}$$

(The number of reactions can be reduced to 20 due to symmetry.)

Note that, in contrast to the fully connected networks (cf. theorem 1), we do not have any convergence guarantee for (8). In fact, in its derivation we made the assumption that certain agents are stochastically independent. While this is a common strategy in the construction of the macroscopic evolution, see for example *mean-field theory* (Porter and Gleeson (2014)), here it introduces an error, cf. Fig. 3. In the next sections, we will employ data-driven techniques to learn the macroscopic evolution and reduce the above mentioned error.

4.2 Learning collective variable dynamics using SINDy

A very popular technique to discover system dynamics from data is called *Sparse Identification of Nonlinear Dynamics* (SINDy) and was proposed by Brunton et al. (2016). It works as follows: Given a time-series of trajectory data $\{z^{(k)}\}$ and a library of basis functions $\{f_j\}$, we numerically approximate the derivatives at the trajectory points $\dot{z}^{(k)}$ (e.g., via finite differences) and solve for the optimal weights $\{\lambda_j\}$ such that $\sum_j \lambda_j f_j(z^{(k)}) \approx \dot{z}^{(k)}$ (linear regression). Additionally, we typically enforce a certain sparsity on the set of weights $\{\lambda_j\}$. Hence, there are many hyperparameters to be considered when applying SINDy: the method to generate derivatives, the library of basis functions, the optimizer and optimizer parameters (especially regarding the desired sparsity of coefficients), etc.

Our data set consist of 9 trajectories of z with different initial conditions, cf. Fig. 2. In order to find optimal hyperparameters, we employ grid search combined with leave-one-out cross-validation, i.e., for every choice of hyperparameters we train 9 models, so that model i trains on all except the i-th trajectory, and define the error as the deviation from the data in infinity norm on the validation trajectory i, averaged over the 9 models. This procedure showed that SINDy is rather sensitive to the choice of hyperparameters. For many choices the dynamics produced by SINDy is unstable, so that we had to artificially bound $z(t)$ to not produce an infinitely large error. The best choice of hyperparameters that we found was using central finite differences for calculating the derivatives, polynomials of degree less than 2 as basis functions, and the SSR optimizer (Boninsegna et al. (2018)). The dynamics produced by SINDy is visualized in Fig. 3 and its error is depicted in Fig. 4. Note that the average error is significantly lower than the error of the analytical model from section 4.1. However, there also exist trajectories where the dynamics learned by SINDy has a bigger error than the analytical model. As discovering the macroscopic dynamics using SINDy is not robust, we will explore an informed learning approach in the next section.

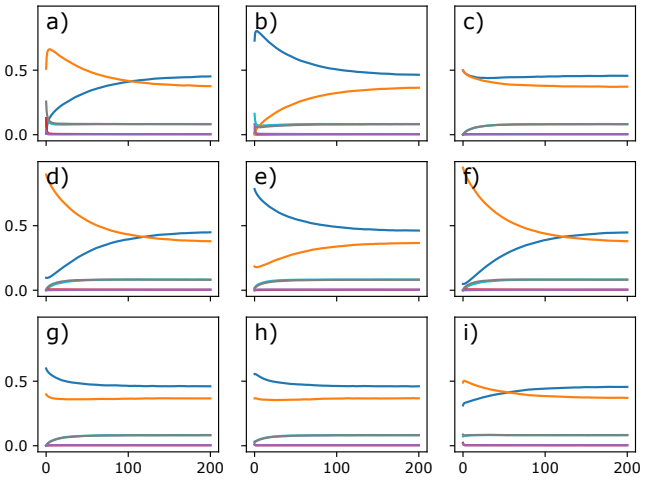


Fig. 2. The available trajectory data consists of 9 trajectories of $\frac{1}{N}\boldsymbol{z}(t)$ (cf. (6)) for $t \in [0, 200]$ and $N = 2000$, starting from different initial conditions and averaged over an ensemble of 500 simulations each. We will refer to the different trajectories by their label, e.g. "trajectory a)".

4.3 Learning the closure

In the analytical derivation of the collective variable equations in section 4.1 we introduced an error by distributing a cumulative rate onto four associated reactions (cf. (7)) under assumptions that may not be correct. Hence, we will employ a data-driven method in order to learn better ways of distributing the shares of the cumulative rate. In contrast to the SINDy technique from the previous section, we do not need to learn the complete dynamics without prior knowledge, but we apply the data-driven method precisely to the unknown part in the analytical equations, thus learning the *closure* of the equations.

To obtain the training data, we iterate through every snapshot of our training trajectories (cf. Fig. 2) and count the occurrences of all possible 5-tuples $(x_{i-2}, \dots, x_{i+2}) \in \{0,1\}^5$, which enables us to calculate the above mentioned shares of the cumulative rates. Then we employ linear

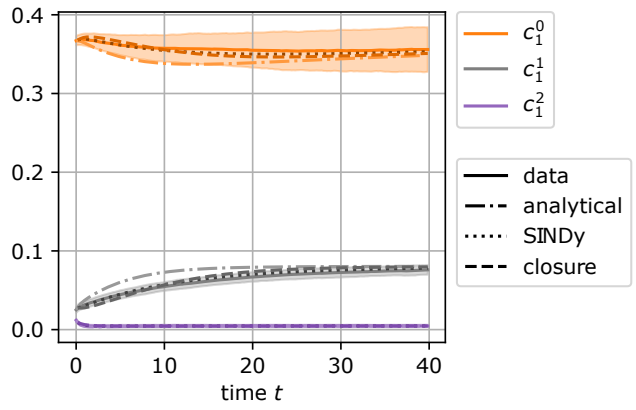


Fig. 3. Trajectory h) data and approximate solutions: "analytical" (cf. section 4.1), "SINDy" (cf. section 4.2), "closure" (cf. section 4.3). $c_m^\ell := y_m^\ell / N$.

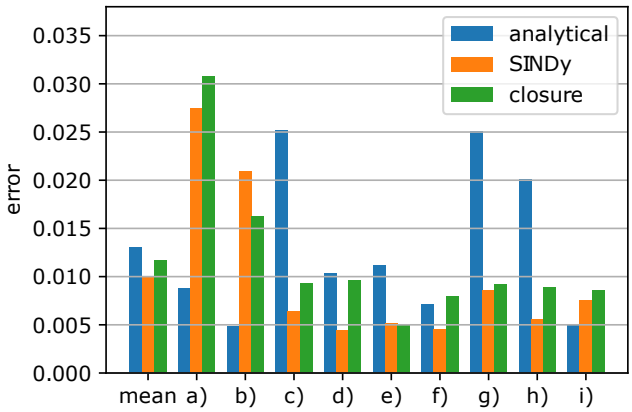


Fig. 4. Error of approximate dynamics for $\boldsymbol{z}$ (cf. (6)) for the 9 test trajectories (cf. Fig. 2): "analytical" (cf. section 4.1), "SINDy" (cf. section 4.2), "closure" (cf. section 4.3). The error is measured as deviation from mean of data w.r.t. infinity norm.

regression with an ℓ_1 regularization, see *LASSO* by Tibshirani (1996), to find an optimal approximation of the empirical share functions within the space of polynomials. We again employ grid search combined with leave-one-out cross-validation (cf. section 4.2) to find the best optimizer parameters and optimal degree of polynomials. As discussed in section 4.1, there are 20 share functions that need to be learned.

We found that this approach of learning the closure is significantly more robust than SINDy (cf. section 4.2) for our example as it rarely produces unstable dynamics, see table 1. However, the error of this method seems to be slightly larger than the error of SINDy, but still lower than the error of the analytical model, cf. Fig. 3 and Fig 4. Note also, that this approach is not inferior to SINDy on all 9 trajectories, e.g., it is superior to SINDy on trajectory b).

Further tests also showed that this approach of learning the closure requires significantly less data than SINDy to produce good results, i.e., we could beat the analytical model using only 5 of the 9 training trajectories while SINDy needed at least 7.

optimizer	library of polynomials of degree		
	1	2	3
STLSQ	33% | 0%	44% | 0%	67% | 0%
SSR	0% | 0%	67% | 0%	100% | 0%

Table 1. Instability of SINDy | closure. The entries show for how many of the 9 training trajectories (cf. Fig. 2) the learned dynamics is unstable. The depicted hyperparameters (optimizer and library) are only a selection of all tested hyperparameters.

5. OUTLOOK

In this extended abstract we showed how one can employ data-driven methods in order to improve the accuracy of conventional macroscopic dynamics approximations. These methods, which we demonstrated for simple network topologies, will be extended to more complex systems in future works. Further tasks are to find the collective variables in a case where we can simulate the system (or some trajectories of it are given), and to derive stochastic models for mesoscopic populations where the deterministic limit of Theorem 1 is not a good approximation.

REFERENCES

Bittracher, A., Koltai, P., Klus, S., Banisch, R., Dellnitz, M., and Schütte, C. (2017). Transition Manifolds of Complex Metastable Systems. *Journal of Nonlinear Science*, 28(2), 471–512. doi:10.1007/s00332-017-9415-0.

Boninsegna, L., Nüske, F., and Clementi, C. (2018). Sparse learning of stochastic dynamical equations. *The Journal of Chemical Physics*, 148(24), 241723. doi:10.1063/1.5018409.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937. doi:10.1073/pnas.1517384113.

Helfmann, L., Conrad, N.D., Djurdjevac, A., Winkelmann, S., and Schütte, C. (2021). From interacting agents to density-based modeling with stochastic PDEs. *Communications in Applied Mathematics and Computational Science*, 16(1), 1–32. doi:10.2140/camcos.2021.16.1.

Kurtz, T.G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and their Applications*, 6(3), 223–240. doi:10.1016/0304-4149(78)90020-0.

Porter, M.A. and Gleeson, J.P. (2014). Dynamical Systems on Networks: A Tutorial.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Winkelmann, S. and Schütte, C. (2020). *Stochastic dynamics in computational biology*. Springer, Cham.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Linear quadratic optimal control of switched differential algebraic equations

Paul Wijnbergen * **Stephan Trenn** *

* *Bernoulli Institute for Maths, CS and AI, University of Groningen, 9747 AG, Groningen, The Netherlands (e-mail: p.wijnbergen@rug.nl, s.trenn@rug.nl).*

Abstract: In this abstract the finite horizon linear quadratic optimal control problem with constraints on the terminal state for switched differential algebraic equations is considered. Furthermore, we seek for an optimal solution that is impulse-free. In order to solve the problem, a non standard finite horizon problem for non-switched DAEs is considered. Necessary and sufficient conditions on the initial value x_0 for solvability of this non standard problem are stated. Based on these results a sequence of subspaces can be defined which lead to necessary and sufficient conditions for solvability of the finite horizon optimal control problem for switched DAEs.

Keywords: Switched systems, Differential Algebraic Equations, Optimal control, Linear systems.

1. INTRODUCTION

In this abstract we consider the following switched differential algebraic system

$$\begin{aligned} E_\sigma \dot{x} &= A_\sigma x + B_\sigma u \\ y &= C_\sigma x + D_\sigma u. \end{aligned} \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{N}$ is the switching signal and $E_p, A_p \in \mathbb{R}^{n\times n}$ and (E_p, A_p) is regular, $B_p \in \mathbb{R}^{n\times m}$, $C_p \in \mathbb{R}^{q\times n}$ and $D_p \in \mathbb{R}^{p\times m}$ for $p, q, n, m \in \mathbb{N}$. We aim to find an impulse-free solution (x, u) on $[t_0, t_f)$ satisfying $x(t_0^-) = x_0$ and $x(t_f^-) \in \mathcal{V}^{\text{end}}$ that minimizes

$$J(x_0, u, t_f) = \int_{t_0}^{t_f} y(t)^\top y(t)\, \mathrm{d}t + x(t_f^-)^\top P x(t_f^-) \tag{2}$$

for some positive semi definite $P = P^\top \in \mathbb{R}^{n\times n}$ and y is the output resulting from the solution (x, u) of (1) satisfying $x(t_0^-) = x_0$. In general, trajectories of switched DAEs exhibit jumps (or even impulses), which may exclude classical solutions from existence. Therefore, we adopt the *piecewise-smooth distributional solution framework* introduced in Trenn (2009).

Differential algebraic equations (DAEs) arise naturally when modeling physical systems with certain algebraic constraints on the state variables. Examples of applications of DAEs in electrical circuits can be found e.g. in Tolsa and Salichs (1993); Riaza (2008); Reis (2010) and gas networks, where the algebraic constraints are induced by the network topology. *e.g.* in Grundel et al. (2014). The algebraic constraints are often eliminated such that the system is described by ordinary differential equations (ODEs). However, in the case of switched systems, the elimination process of the constraints is in general different for each individual mode. Therefore, in general, there does not exist a description as a switched ODE with a common state variable for every mode. This problem can be overcome by studying switched DAEs directly.

The literature on optimal control of non-switched DAEs is quite mature, (besides the already mentioned literature) see for the finite horizon e.g. Kunkel and Mehrmann (2008, 1997); Ilchmann et al. (2019, 2021); Wijnbergen and Trenn (2021b) on a finite horizon and for the infinite time horizon see e.g. Cobb (1983); Mehrmann (1989); Reis et al. (2015); Reis and Voigt (2019); Bankmann and Voigt (2019). Furthermore, several structural properties of switched DAEs have been investigated recently (Wijnbergen and Trenn, 2021a, 2020). However, to the best of the authors knowledge, optimal control of switched DAEs has not been studied yet.

The finite horizon problem is motivated by the study of optimal control on an infinite horizon, *i.e.,* the minimization of

$$J(x_0, u) = \int_{t_0}^{\infty} y(t)^\top y(t)\, \mathrm{d}t. \tag{3}$$

on the interval $[t_0, \infty)$ while using a dynamic programming approach. It can be shown that if there exists an input that minimizes (3), the optimal cost resulting from the interval $[t_f, \infty)$ is quadratic in $x(t_f^-)$, *i.e.,* $x(t_f^-)^\top P x(t_f^-)$ for some matrix $P \in \mathbb{R}^{n\times n}$. This result allows for a dynamic programming approach. Assuming that the matrix P and the optimal input u restricted to $[t_f, \infty)$ are known, it follows that the minimization of (3) is equivalent to finding the input u restricted to $[t_0, t_f)$ such that

$$\begin{aligned} J(x_0, u) &= \int_{t_0}^{t_f} y(t)^\top y(t)\, \mathrm{d}t + \int_{t_f}^{\infty} y(t)^\top y(t)\, \mathrm{d}t \\ &= \int_{t_0}^{t_f} y(t)^\top y(t)\, \mathrm{d}t + x(t_f^-)^\top P x(t_f^-). \end{aligned}$$

is minimal.

⋆ This work was supported by the NWO Vidi-grant 639.032.733.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

For many real word applications Dirac impulses in the state are to be avoided as they can cause damage to components of the system or create hazardous situations. Therefore, we aim to find an optimal *impulse-free* solution (x, u). However, there generally only exists an optimal (impulse-free) solution on $[t_f, \infty)$ if the state at t_f is contained in some subspace, *i.e.,* $x(t_f^-) \in \mathcal{V}^{\text{end}}$ for some subspace $\mathcal{V}^{\text{end}} \subseteq \mathbb{R}^n$.

In order to solve the problem for switched systems with a switching signal that induces an arbitrary yet finitely many modes on the interval $[t_0, t_f)$ we will first consider the case that only two modes are induced on $[t_0, \infty)$ and the switch occurs at t_f. Within this context we aim to minimize (2) with respect to a non-switched DAE. Once conditions for the single switched case are obtained, conditions for the general case will follow straightforwardly. Hence first we will focus on finding an optimal solution to

$$E\dot{x} = Ax + Bu, \tag{4}$$
$$y = Cx + Du \tag{5}$$

that minimizes (2) under the constraint $x_0 \in \mathbb{R}^n$ and $x(t_f^-) \in \mathcal{V}^{\text{end}}$.

As the terminal cost matrix P represents the cost resulting from the interval $[t_f, \infty)$ and the mode active on this interval is not necessarily structurally related to the dynamics (4), we can only assume that $P \in \mathbb{R}^{n\times n}$ is some positive semi-definite matrix. This is in contrast to the assumption commonly made in the literature for optimal control of non switched DAEs that the terminal cost matrix is of the form $P = E^\top \tilde{P} E$ for some positive semi-definite $\tilde{P} \in \mathbb{R}^{n\times n}$ (Lewis, 1985; Bender and Laub, 1985; Katayama and Minamino, 1992). Also note that whereas commonly a closed interval is of interest, in this paper a half open interval is considered. Consequently, the terminal cost can penalizes algebraic states and as a result $x(t_f^-)$ is not necessarily equal to $x(t_f)$ or even well defined such that an optimal solution might fail to exist.

The remainder of this paper is structured as follows. The mathematical preliminaries and the main results are given in Section 2 and 3, respectively. Conclusions and a discussion on future work are given in Section 4.

2. MATHEMATICAL PRELIMINARIES

In the following, we consider *regular* matrix pairs (E, A), i.e. for which the polynomial $\det(sE - A)$ is not the zero polynomial. Recall the following result on the *quasi-Weierstrass form (QWF)* (Berger et al., 2012).

Proposition 1. A matrix pair $(E, A) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ is regular if, and only if, there exists invertible matrices $S, T \in \mathbb{R}^{n\times n}$ such that

$$(SET, SAT) = \left(\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} \right), \tag{6}$$

where $J \in \mathbb{R}^{n_1\times n_1}$, $0 \leqslant n_1 \leqslant n$, is some matrix and $N \in \mathbb{R}^{n_2\times n_2}$, $n_2 := n - n_1$, is a nilpotent matrix.

The matrices S and T can be calculated by using the so-called *Wong sequences* (Berger et al., 2012; Wong, 1974): Based on the Wong sequences we define the following projectors and selectors.

Definition 2. Consider the regular matrix pair (E, A) with corresponding quasi-Weierstrass form (6). The *consistency projector* of (E, A) is given by

$$\Pi := T \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} T^{-1},$$

the *differential selector* and the *impulse selector* are given by

$$\Pi^{\text{diff}} := T \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} S, \quad \Pi^{\text{imp}} := T \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} S.$$

respectively

In all three cases the block structure corresponds to the block structure of the QWF. Furthermore we define

$$A^{\text{diff}} := \Pi^{\text{diff}} A, \qquad E^{\text{imp}} := \Pi^{\text{imp}} E,$$
$$B^{\text{diff}} := \Pi^{\text{diff}} B, \qquad B^{\text{imp}} := \Pi^{\text{imp}} B.$$

Note that all the above defined matrices do not depend on the specifically chosen transformation matrices S and T; they are uniquely determined by the original regular matrix pair (E, A). An important feature for DAEs is the so called consistency space, defined as follows for the DAE given in (4).

Definition 3. Consider the DAE (4), then the *consistency space* is defined as

$$\mathcal{V}_{(E,A)} := \left\{ x_0 \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \exists \text{ smooth solution } x \text{ of (4)} \\ \text{with } u = 0 \text{ and } x(0) = x_0 \end{array} \right\},$$

and the *augmented consistency space* is defined as

$$\mathcal{V}_{(E,A,B)} := \left\{ x_0 \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \exists \text{ smooth solutions } (x, u) \text{ of (4)} \\ \text{with } x(0) = x_0 \end{array} \right\}.$$

For studying impulsive solutions of (4), we consider the space of *piecewise-smooth distributions* $\mathbb{D}_{\text{pw}\mathcal{C}^\infty}$ from Trenn (2009) as the solution space. That is, we seek a solution $(x, u) \in (\mathbb{D}_{\text{pw}\mathcal{C}^\infty})^{n+m}$ to the following initial-trajectory problem (ITP) associated with (4):

$$x_{(-\infty,0)} = x^0_{(-\infty,0)}, \tag{7a}$$
$$(E\dot{x})_{[0,\infty)} = (Ax)_{[0,\infty)} + (Bu)_{[0,\infty)}, \tag{7b}$$

where $x^0 \in (\mathbb{D}_{\text{pw}\mathcal{C}^\infty})^n$ is some initial trajectory, and $f_{\mathcal{I}}$ denotes the restriction of a piecewise-smooth distribution f to an interval $\mathcal{I}$. In Trenn (2009) it is shown that the ITP (7) has a unique solution for any initial trajectory if, and only if, the matrix pair (E, A) is regular. As a direct consequence, the switched DAE considered in (1) with regular matrix pairs is also uniquely solvable (with piecewise-smooth distributional solutions) for any switching signal with locally finitely many switches.

The space of initial values for which there exists an impulse-free solution (x, u) of (4) is defined as follows.

Definition 4. Consider the DAE (4), then the *impulse controllable space* is defined as

$$\mathcal{C}^{\text{imp}} := \left\{ x_0 \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \exists \text{ solution of (7) satisfying} \\ x(0^-) = x_0 \text{ and } (x, u)[0] = 0 \end{array} \right\},$$

The DAE is called impulse controllable if $\mathcal{C}^{\text{imp}} = \mathbb{R}^n$.

Lemma 5. (Cobb (1981)). The DAE (4) is impulse controllable if and only if there exist a feedback input $u = Lx + v$ such that in terms of the selectors Π^{diff} and Π^{imp} resulting from the Wong sequences based on $(E, A + BL)$, the solution x is given by $x = x^{\text{diff}} - B^{\text{imp}} v$ where x^{diff} solves

$$\dot{x}^{\mathrm{diff}} = (A + BL)^{\mathrm{diff}} x^{\mathrm{diff}} + B^{\mathrm{diff}} v \tag{8}$$

Given a matrix pair (E, A) and a matrix B we can always generate an impulse-controllable DAE with the same solution behavior for initial values $x_0 \in \mathcal{C}^{\mathrm{imp}}$.

Lemma 6. Consider the DAE (4). A solution (x, u) satisfying $x(0^-) = x_0 \in \mathcal{C}^{\mathrm{imp}}$ solves (4) if and only if (x, u) solves

$$EW\dot{x} = Ax + Bu \tag{9}$$

where W is an orthogonal projector onto $\mathcal{C}^{\mathrm{imp}}$. In addition, (9) is impulse controllable.

3. MAIN RESULTS

The concepts introduced in the previous section are now utilized to study the minimization of (2) subject to the non-switched DAE (4). We aim to minimize (2) over all (x, u) that are impulse-free and satisfy $x(t_0^-) = x_0$ and $x(t_f^-) \in \mathcal{V}^{\mathrm{end}}$.

As we aim to find an impulse-free solution (x, u) that minimizes (2), it is necessary that $x(t_0) = x_0 \in \mathcal{C}^{\mathrm{imp}}$. Consequently, it follows from Lemma 6 that we can assume without loss of generality that (4) is impulse controllable. Moreover, we can assume that an index-reducing feedback in the sense of Lemma 5 has been applied and that the system is of index-1.

In the case (4) is of index-1, we observe, by making use of the decomposition $x = x^{\mathrm{diff}} - B^{\mathrm{imp}}u$, that (x, u, y) solves (4) if, and only if, $(x^{\mathrm{diff}}, u, y)$ with $x^{\mathrm{diff}}(t_0^-) = \Pi x_0$ solves

$$\begin{aligned} \dot{x}^{\mathrm{diff}} &= A^{\mathrm{diff}} x^{\mathrm{diff}} + B^{\mathrm{diff}} u \\ \bar{y} &= C x^{\mathrm{diff}} + (D - CB^{\mathrm{imp}})u \end{aligned} \tag{10}$$

which shows that the minmization of (2) subject to (4) is equivalent to the minimization of

$$\bar{J}(x^{\mathrm{diff}}, u) = \int_{t_0}^{t_f} \bar{y}(t)^\top \bar{y}(t)\, \mathrm{d}t + \bar{x}(t_f^-)^\top P \bar{x}(t_f). \tag{11}$$

where $\bar{x} := x^{\mathrm{diff}} - B^{\mathrm{imp}}u$, subject to (10). This shows that the optimal control problem for non switched DAEs can be reduced to an equivalent problem for ordinary differential equations (ODEs). However, note that the latter problem is still not a standard finite horizon linear quadratic optimal control problem for ODEs as the terminal state of the input is penalized by the terminal cost and because of the subspace endpoint constraint $\bar{x}(t_f^-) = x(t_f^-) \in \mathcal{V}^{\mathrm{end}}$.

In order to ensure that the optimal input does not contain impulses, we assume that $\bar{D} := D - CB^{\mathrm{imp}}$ has full column rank, such that $\bar{D}^\top \bar{D}$ is positive definite. For ODE optimal control problems with a cost resulting from an output (10) this is standard. However, in the literature on optimal control on DAE it is often assumed that $D^\top D$ is positive definite, which we do not require here. Note that a sufficient condition for $\bar{D}$ to have full column rank is

$$\operatorname{rank} [CW \;\; D] = m$$

were W is some projector onto $\ker E$. This assumption is very similar to the assumption that the system (4) is impulse-observable.

3.1 Regarding the terminal cost

As the terminal cost in (11) penalizes the input, it follows that if there exists a solution that minimizes (11), the value of $u(t_f^-)$ must be well defined and satisfies given the optimal state $x^{\mathrm{diff}}(t_f^-)$

$$\begin{aligned} &((x^{\mathrm{diff}}(t_f^-) - B^{\mathrm{imp}}u(t_f^-))^\top P((x^{\mathrm{diff}}(t_f^-) - B^{\mathrm{imp}}u(t_f^-)) \\ &\quad \leqslant ((x^{\mathrm{diff}}(t_f^-) - B^{\mathrm{imp}}v)^\top P((x^{\mathrm{diff}}(t_f^-) - B^{\mathrm{imp}}v)) \end{aligned}$$

for all v satisfying $x^{\mathrm{diff}}(t_f^-) - B^{\mathrm{imp}}v \in \mathcal{V}^{\mathrm{end}}$. Using Lagrange multipliers and noting that the terminal cost is a convex function of $u(t_f^-)$ leads to the following result.

Lemma 7. Let (x^{diff}, u) be a solution satisfying $\bar{x}(t_0^-) = x_0$ and $\bar{x}(t_f^-) \in \mathcal{V}^{\mathrm{end}}$ that minimizes (11). Then the terminal cost satisfies

$$\bar{x}(t_f^-)^\top P \bar{x}(t_f^-) = x^{\mathrm{diff}}(t_f^-)^\top \Psi^\top P \Psi x^{\mathrm{diff}}(t_f^-). \tag{12}$$

where $\Psi = (I - B^{\mathrm{imp}}N)$ for some N satisfying

$$\begin{bmatrix} I & 0 & N \end{bmatrix} \ker \begin{bmatrix} B^{\mathrm{imp}\top} P B^{\mathrm{imp}} & B^{\mathrm{imp}\top} \Pi_{\mathcal{V}^\perp}^\top & -2B^{\mathrm{imp}\top} P \Pi \\ \Pi_{\mathcal{V}^\perp} B^{\mathrm{imp}} & 0 & -\Pi_{\mathcal{V}^\perp} \Pi \end{bmatrix} = 0.$$

where $\Pi_{\mathcal{V}^\perp}$ is an orthogonal projector onto the orthogonal complement of $\mathcal{V}^{\mathrm{end}}$.

It follows from Lemma 7 that instead of minimizing (11) directly, we can focus on finding an input that minimizes

$$\begin{aligned} \bar{J}_\Psi(x^{\mathrm{diff}}, u) = \int_{t_0}^{t_f} \bar{y}(t)\bar{y}(t)\, \mathrm{d}t \\ + x^{\mathrm{diff}}(t_f^-)^\top \Psi^\top P \Psi x^{\mathrm{diff}}(t_f^-) \end{aligned} \tag{13}$$

and verify whether the optimal input satisfies (12). However, the computation of the input that minimizes (13) is rather straightforward. After denoting

$$\bar{y}(t)^\top \bar{y}(t) = \begin{bmatrix} x^{\mathrm{diff}} \\ u \end{bmatrix}^\top \begin{bmatrix} Q & S^\top \\ S & R \end{bmatrix} \begin{bmatrix} x^{\mathrm{diff}} \\ u \end{bmatrix}$$

we can state the following result.

Lemma 8. The cost functional $\bar{J}_\Psi(x^{\mathrm{diff}}, u)$ satisfies

$$\begin{aligned} &\bar{J}_\Psi(x^{\mathrm{diff}}, u) - x^{\mathrm{diff}}(t_0^-)^\top X(t_0) x^{\mathrm{diff}}(t_0^-) \\ &\qquad = \int_{t_0}^{t_f} \left(\left\| u + R^{-1}(\bar{B}^{\mathrm{diff}\top} X + S^\top) \bar{x}^{\mathrm{diff}} \right\|_2^2 \right) \end{aligned}$$

where X solves

$$\begin{aligned} \dot{X} = A^{\mathrm{diff}\top} X + X^\top A^{\mathrm{diff}} + Q \\ - (S + X^\top B^{\mathrm{diff}}) R^{-1} (B^{\mathrm{diff}\top} X + S^\top). \end{aligned} \tag{14}$$

with terminal condition $X(t_f) = \Psi^\top P \Psi$.

Corollary 9. If an input u minimizes (13), then

$$u = -R^{-1}(B^{\mathrm{diff}\top} X + S^\top) x^{\mathrm{diff}} \tag{15}$$

where X is a solution to (14) with $X(t_f) = \Psi^\top P \Psi$.

The result of Corollary 9 shows that if there exists an optimal control, it needs to be of a particular form. However, a feedback of the form (15) does not necessarily controls an initial value Πx_0 to the desired subspace, *e.g.,* in the case $\mathcal{V}^{\mathrm{end}}$ is the zero subspace, and hence an optimal control might fail to exist. In order to determine which intial values are controlled to $\mathcal{V}^{\mathrm{end}}$ at t_f^- we define the following flow operator.

Definition 10. The backwards state transition matrix for the closed loop ODE

$$\dot{x}^{\text{diff}} = \left(A^{\text{diff}} - B^{\text{diff}} R^{-1} \left(B^{\text{diff}\top} X + S^\top\right)\right) x^{\text{diff}}$$

is given by $\Omega(t, t_f)$. Hence $x^{\text{diff}}(t) = \Omega(t, t_f) x^{\text{diff}}(t_f^-)$.

Recall that the state $x = x^{\text{diff}} - B^{\text{imp}} u$ and thus for the input (15) we have $x(t_f^-) = M x^{\text{diff}}(t_f^-)$ where

$$M := I - B^{\text{imp}} R^{-1} \left(B^{\text{diff}\top} \Psi^\top P \Psi + S^\top\right).$$

As $x^{\text{diff}}(t_f^-) = \Pi\xi$ for some $\xi \in \mathbb{R}^n$ it follows that $x(t_f^-) \in \mathcal{V}^{\text{end}}$ if and only if

$$x^{\text{diff}}(t_f^-) = \Pi\xi \in \ker \Pi_{\mathcal{V}^\perp} M$$

Next, observe that the input (15) satisfies (12) if and only if

$$\begin{aligned} x(t_f^-)^\top P x(t_f^-) &= x^{\text{diff}}(t_f^-)^\top M^\top P M \Pi x(t_f^-) \\ &= x^{\text{diff}}(t_f^-)^\top \Psi^\top P \Psi x^{\text{diff}}(t_f^-). \end{aligned}$$

This is the case if $M^\top P M x^{\text{diff}}(t_f^-) = \Psi^\top P \Psi x^{\text{diff}}(t_f^-)$. Given these observations, we can state the following result regarding the minimization of (11).

Theorem 11. There exists an impulse-free solution (x, u) satisfying $x(t_0^-) = x_0$ and $x(t_f^-) \in \mathcal{V}^{\text{end}}$ that minimizes (11) if and only if

$$x_0 \in \mathcal{V}^{\text{init}} := \Omega(t_0, t_f) \ker \begin{bmatrix} \Pi_{\mathcal{V}^\perp} M \\ M^\top P M - \Psi^\top P \Psi \end{bmatrix} \Pi.$$

3.2 Multiple switched case

Given the result of Theorem 11 we are now able to state conditions for the existence of a solution that minimizes (2) subject to (1). To do so, we define the following sequence

$$\begin{aligned} \mathcal{V}_{\mathtt{n}}^{\text{end}} &= \mathcal{V}^{\text{end}}, \\ \mathcal{V}_{i-1}^{\text{end}} &= \mathcal{V}_i^{\text{init}}, \end{aligned} \qquad i = \mathtt{n}, \mathtt{n}-1, ..., 0$$

where $\mathcal{V}_i^{\text{init}}$ is defined according to Theorem 11 on the interval $[t_i, t_{i+1})$ w.r.t. $\mathcal{V}_i^{\text{end}}$.

Theorem 12. There exists an impulse-free solution (x, u) that minizes (2) and satisfies $x(t_0^-) = x_0$ and $x(t_f^-) \in \mathcal{V}^{\text{end}}$ if and only if $x_0 \in \mathcal{V}_0^{\text{init}}$.

4. CONCLUSION

In this abstract we considered the finite horizon optimal control problem for switched DAEs. Based on sovability of `n` nonstandard optimal control problems for ODEs solvability of the optimal control problem for switched DAEs can be concluded. In this abstract impulse-freeness of the solution (x, u) was required. A natural direction of research is to investigate the existence of optimal impulsive solutions.

REFERENCES

Bankmann, D. and Voigt, M. (2019). On linear-quadratic optimal control of implicit difference equations. *IMA Journal of Mathematical Control and Information*, 36(3), 779–833.

Bender, D.J. and Laub, A.J. (1985). The linear-quadratic optimal regulator for descriptor systems. In *Proc. 24th IEEE Conf. Decis. Control, Ft. Lauderdale, FL*, 957–962. doi:10.1109/CDC.1985.268642.

Berger, T., Ilchmann, A., and Trenn, S. (2012). The quasi-Weierstraß form for regular matrix pencils. *Linear Algebra Appl.*, 436(10), 4052–4069. doi:10.1016/j.laa.2009.12.036.

Cobb, J.D. (1981). Feedback and pole placement in descriptor variable systems. *Int. J. Control*, 33(6), 1135–1146.

Cobb, J.D. (1983). Descriptor variable systems and optimal state regulation. *IEEE Trans. Autom. Control*, 28, 601–611. doi: 10.1109/TAC.1983.1103283.

Grundel, S., Jansen, L., Hornung, N., Clees, T., Tischendorf, C., and Benner, P. (2014). Model order reduction of differential algebraic equations arising from the simulation of gas transport networks. In *Progress in differential-algebraic equations*, 183–205. Springer.

Ilchmann, A., Leben, L., Witschel, J., and Worthmann, K. (2019). Optimal control of differential-algebraic equations from an ordinary differential equation perspective. *Optimal Control Applications and Methods*, 40(2), 351–366.

Ilchmann, A., Witschel, J., and Worthmann, K. (2021). Model predictive control for singular differential-algebraic equations. *International Journal of Control*, 1–10.

Katayama, T. and Minamino, K. (1992). Linear quadratic regulator and spectral factorization for continous-time descriptor systems. In *[1992] Proceedings of the 31st IEEE Conference on Decision and Control*, 967–972. IEEE.

Kunkel, P. and Mehrmann, V. (1997). The linear quadratic control problem for linear descriptor systems with variable coefficients. *Math. Control Signals Syst.*, 10, 247–264.

Kunkel, P. and Mehrmann, V. (2008). Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index. *Math. Control Signals Syst.*, 20, 227–269.

Lewis, F.L. (1985). Optimal control for singular systems. In *1985 24th IEEE Conference on Decision and Control*, 266–272. IEEE.

Mehrmann, V. (1989). Existence, uniqueness and stability of solutions to singular, linear-quadratic control problems. *Linear Algebra Appl.*, 291–331.

Reis, T. (2010). Circuit synthesis of passive descriptor systems - a modified nodal approach. *Int. J. Circ. Theor. Appl.*, 38, 44–68.

Reis, T., Rendel, O., and Voigt, M. (2015). The Kalman-Yakubovich-Popov inequality for differential-algebraic systems. *Linear Algebra Appl.*, 86, 153–193.

Reis, T. and Voigt, M. (2019). Linear-quadratic optimal control of differential-algebraic systems: the infinite time horizon problem with zero terminal state. *SIAM Journal on Control and Optimization*, 57(3), 1567–1596.

Riaza, R. (2008). *Differential-Algebraic Systems. Analytical Aspects and Circuit Applications*. World Scientific Publishing, Basel.

Tolsa, J. and Salichs, M. (1993). Analysis of linear networks with inconsistent initial conditions. *IEEE Trans. Circuits Syst.*, 40(12), 885 – 894. doi:10.1109/81.269029.

Trenn, S. (2009). *Distributional differential algebraic equations*. Ph.D. thesis, Institut für Mathematik, Technische Universität Ilmenau, Universitätsverlag Ilmenau, Germany. URL `http://www.db-thueringen.de/servlets/DocumentServlet?id=13581`.

Wijnbergen, P. and Trenn, S. (2020). Impulse controllability of switched differential-algebraic equations. In *2020 European Control Conference (ECC)*, 1561–1566. IEEE.

Wijnbergen, P. and Trenn, S. (2021a). Impulse-free interval-stabilization of switched differential algebraic equations. *Systems & Control Letters*, 149, 104870.

Wijnbergen, P. and Trenn, S. (2021b). Optimal control of daes with unconstrained terminal costs. In *2021 60th IEEE Conference on Decision and Control (CDC)*, 5275–5280. IEEE.

Wong, K.T. (1974). The eigenvalue problem $\lambda Tx + Sx$. *J. Diff. Eqns.*, 16, 270–280. doi:10.1016/0022-0396(74)90014-X.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On neural network architectures for solving high dimensional Hamilton-Jacobi equations arising in optimal control

Jérôme Darbon * **, Peter M. Dower** ** **, Tingwei Meng** ***

* *Division of Applied Mathematics, Brown University (e-mail: jerome_darbon@brown.edu)*
** *Department of Electrical and Electronic Engineering, The University of Melbourne (e-mail: pdower@unimelb.edu.au)*
*** *Department of mathematics, UCLA, Los Angeles, CA, USA (e-mail: tingwei@math.ucla.edu)*

Abstract: We propose new mathematical connections between Hamilton-Jacobi (HJ) partial differential equations (PDEs) with initial data and neural network architectures. Specifically, we prove that some classes of neural networks correspond to representation formulas of HJ PDE solutions whose Hamiltonians and initial data are obtained from the parameters or the activation functions of the neural networks. These results do not require any learning stage. In addition these results do not rely on universal approximation properties of neural networks; rather, our results show that some classes of neural network architectures naturally encode the physics contained in some HJ PDEs. Our results naturally yield efficient neural network-based methods for evaluating solutions of some HJ PDEs in high dimension without using grids or numerical approximations.

Keywords: Optimal Control, Hamilton-Jacobi Partial Differential Equations, Neural Networks, Grid-Free Numerical Methods, High Dimensions

1. INTRODUCTION

Hamilton-Jacobi (HJ) partial differential equations (PDEs) are widely used in physics, optimal control, game theory, and imaging sciences. An HJ PDE is given as follows

$$\begin{cases} \dfrac{\partial S}{\partial t}(\boldsymbol{x},t) + H(t,\boldsymbol{x},\nabla_{\boldsymbol{x}} S(\boldsymbol{x},t)) = 0 & \text{in } \mathbb{R}^n \times (0,+\infty), \\ S(\boldsymbol{x},0) = J(\boldsymbol{x}) & \text{in } \mathbb{R}^n, \end{cases} \tag{1}$$

where $S\colon \mathbb{R}^n \times [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$ denotes the solution to the PDE, $H\colon [0,+\infty) \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is the Hamiltonian, and $J\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is the initial condition. The computational complexity of standard grid-based numerical algorithms for solving HJ PDEs scales exponentially with respect to the dimension n. This exponential scaling is often referred to as the "curse of dimensionality" (CoD) [Bellman (1961)]. Due to the CoD, these grid-based methods are infeasible for solving high-dimensional problems (e.g., dimensions greater than five) and thus are infeasible for many practical applications.

In the literature, several methods are proposed to overcome the curse of dimensionality when solving high dimensional HJ PDEs and optimal control problems. These methods include, but are not limited to, max-plus methods [Akian et al. (2008); Fleming and McEneaney (2000); McEneaney (2006)], optimization methods [Darbon and Osher (2016); Yegorov and Dower (2017)], tensor decomposition techniques [Dolgov et al. (2019); Horowitz et al. (2014); Todorov (2009)], sparse grids [Bokanowski et al. (2013); Kang and Wilcox (2017)], polynomial approximation [Kalise et al. (2019)], model order reduction [Alla et al. (2017); Kunisch et al. (2004)], dynamic programming and reinforcement learning [Alla et al. (2019); Bertsekas (2019)] and neural networks [Bachouch et al. (2018); Jiang et al. (2016); Nakamura-Zimmerer et al. (2019); Jin et al. (2020)].

Recently, we proposed several neural network architectures for solving different classes of high-dimensional HJ PDEs [Darbon et al. (2020); Darbon and Meng (2021); Darbon et al. (2021)]. These neural network architectures have solid theoretical guarantees from the theory of HJ PDEs, and they can leverage efficient hardware dedicated to neural networks and designed for future real-time applications. By using the theory of HJ PDEs, the parameters in these neural networks are assigned directly from the PDEs. Therefore, we do not need any training process. This is the main difference of our proposed neural network methods with traditional neural network methods.

In the following sections, we summarize our proposed three architectures. The first architecture is a shallow neural network proposed in Darbon et al. (2020) and is summarized in Section 2. The second is a neural network with two hidden layers proposed in Darbon and Meng (2021) and is summarized in Section 3. The last one is a deep neural network proposed in Darbon et al. (2021) and is summarized in Section 4.

⋆ This research is supported by NSF 1820821 and AFOSR MURI FA9550-20-1-0358.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Throughout, we use $\mathbb{R}^{n\times l}$ to denote the set of matrices with n rows and l columns with entries in $\mathbb{R}$, and use S^n to denote the set of real-valued symmetric matrices in $\mathbb{R}^{n\times n}$. We use $C(0,T;X)$ to denote the set of continuous functions from $[0,T]$ to a space X.

2. A SHALLOW NEURAL NETWORK ARCHITECTURE

A shallow neural network architecture for solving a class of high-dimensional HJ PDEs is proposed in Darbon et al. (2020). The target PDE is a subclass of HJ PDEs (1) whose Hamiltonian H does not depend on $(\boldsymbol{x},t)$, and whose initial condition J is convex. We define the neural network function $f\colon \mathbb{R}^n\times[0,+\infty)\to\mathbb{R}$ by

$$f(\boldsymbol{x},t) \doteq \max_{i\in\{1,\dots,m\}} \{\langle \boldsymbol{p}_i,\boldsymbol{x}\rangle - t\theta_i - \gamma_i\}, \tag{2}$$

with parameters $\boldsymbol{p}_i\in\mathbb{R}^n$, $\theta_i\in\mathbb{R}$, and $\gamma_i\in\mathbb{R}$ for $i=1,\dots,m$ (where i is the index for the neuron, and m is the total number of neurons involved). The illustration for this neural network architecture is shown in Fig. 1.

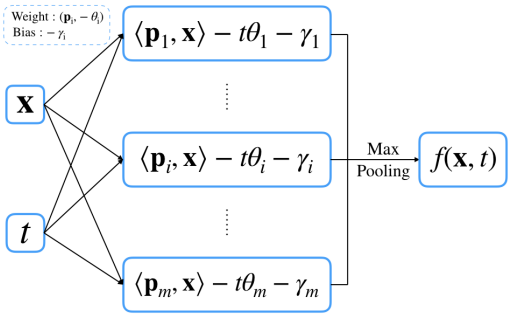


Fig. 1. Architecture of the neural network (2) that represents the viscosity solution to certain class of HJ PDEs, whose Hamiltonian does not depend on $(\boldsymbol{x},t)$, and whose initial condition is convex.

In Darbon et al. (2020), both theoretical guarantees and numerical experiments are provided for this architecture. Under certain assumptions, the neural network function f computes the viscosity solution to the HJ PDE (1) whose initial condition $J\colon\mathbb{R}^n\to\mathbb{R}$ and Hamiltonian $H\colon\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ are defined using the parameters $\{(\boldsymbol{p}_i,\theta_i,\gamma_i)\}_{i=1}^m$ as follows

$$J(\boldsymbol{x}) \doteq \max_{i\in\{1,\dots,m\}} \{\langle \boldsymbol{p}_i,\boldsymbol{x}\rangle - \gamma_i\}, \tag{3}$$

and

$$H(\boldsymbol{p}) \doteq \begin{cases} \inf\limits_{\boldsymbol{\alpha}\in\mathcal{A}(\boldsymbol{p})} \left\{\sum\limits_{i=1}^m \alpha_i\theta_i\right\}, & \text{if } \boldsymbol{p}\in \operatorname{dom} J^*,\\ +\infty, & \text{otherwise,}\end{cases} \tag{4}$$

where J^* denotes the Fenchel-Legendre transform of the function J (see Hiriart-Urruty and Lemaréchal (1993b)), and the set $\mathcal{A}(\boldsymbol{p})\subseteq\mathbb{R}^m$ is defined by

$$\mathcal{A}(\boldsymbol{p}) \doteq \operatorname*{arg\,min}_{\substack{(\alpha_1,\dots\alpha_m)\in\Lambda_m\\ \sum_{i=1}^m \alpha_i\boldsymbol{p}_i=\boldsymbol{p}}} \left\{\sum_{i=1}^m \alpha_i\gamma_i\right\}.$$

Here, we use Λ_m to denote the unit simplex in $\mathbb{R}^m$, i.e.,

$$\Lambda_m \doteq \left\{(\alpha_1,\dots,\alpha_m)\in[0,1]^m : \sum_{i=1}^m \alpha_i = 1\right\}.$$

We also proved that the function f with the same parameters $\{(\boldsymbol{p}_i,\theta_i,\gamma_i)\}_{i=1}^m$ may solve different HJ PDEs. The function H defined in (4) is a lower bound for all possible Hamiltonians in these HJ PDEs. In this way, we identified the smallest Hamiltonian H in (4) for the set of HJ PDEs whose viscosity solution is the function f with certain parameters $\{(\boldsymbol{p}_i,\theta_i,\gamma_i)\}_{i=1}^m$. We refer readers to Darbon et al. (2020) for detailed assumptions and characterization for this set of HJ PDEs. In Darbon et al. (2020), we also provided several numerical experiments showing the ability of our proposed architectures for overcoming the CoD.

3. A NEURAL NETWORK ARCHITECTURE WITH TWO HIDDEN LAYERS

A neural network architecture, with two hidden layers, for solving another class of high-dimensional HJ PDEs is proposed in Darbon and Meng (2021). This class assumes a convex Hamiltonian that is independent of (x,t). The neural network function $f\colon\mathbb{R}^n\times[0,+\infty)\to\mathbb{R}$ involved is defined by

$$f(\boldsymbol{x},t) \doteq \min_{i\in\{1,\dots,m\}} \left\{tL\left(\frac{\boldsymbol{x}-\boldsymbol{u}_i}{t}\right) + a_i\right\}, \tag{5}$$

with parameters $\boldsymbol{u}_i\in\mathbb{R}^n$ and $a_i\in\mathbb{R}$ for $i=1,\dots,m$. The activation function $L\colon\mathbb{R}^n\to\mathbb{R}$ corresponds to the Lagrangian function in the theory of HJ PDEs and optimal control problems. An illustration is shown in Fig. 2.

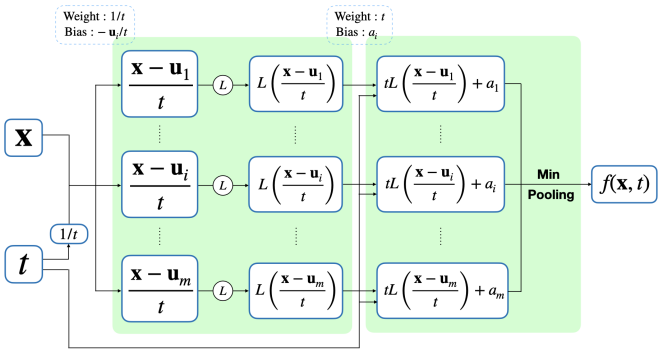


Fig. 2. Architecture of the neural network (5) that represents the viscosity solution to certain class of HJ PDEs whose Hamiltonian is convex and does not depend on $(\boldsymbol{x},t)$.

The initial data $J\colon\mathbb{R}^n\to\mathbb{R}$ is defined by

$$J(\boldsymbol{x}) \doteq \min_{i\in\{1,\dots,m\}} \{L'_\infty(\boldsymbol{x}-\boldsymbol{u}_i) + a_i\}, \tag{6}$$

where L'_∞ is the asymptotic function of L (see (Hiriart-Urruty and Lemaréchal, 1993a, Chap. IV.3.2)). Under certain assumptions, the function f defined in (5) is shown to compute the viscosity solution to the HJ PDE (1) with Hamiltonian $H=L^*$ (i.e., H equals the Fenchel-Legendre transform of the activation function L in the neural network, and H does not depend on $(\boldsymbol{x},t)$) and initial condition J defined in (6). Although the initial

condition J has a specific form in (6), this form can be used to approximate other meaningful initial conditions when m approaches infinity. In Darbon and Meng (2021), we also provided several numerical experiments in high dimensions. From the experimental results, we observe that our proposed architecture can overcome the CoD for certain HJ PDEs. Another neural network architecture with two hidden layers is proposed in the same paper. We refer readers to Darbon and Meng (2021) for more details about the theoretical guarantees and numerical results for these two architectures.

4. A DEEP NEURAL NETWORK ARCHITECTURE

A deep neural network architecture is proposed in our recent paper Darbon et al. (2021) for representing the viscosity solution to the backward HJ PDE

$$\begin{cases} -\dfrac{\partial V(\boldsymbol{x},t)}{\partial t} + H(t,\boldsymbol{x},\nabla_{\boldsymbol{x}} V(\boldsymbol{x},t)) = 0 & \boldsymbol{x}\in\mathbb{R}^n, t\in(0,T), \\ V(\boldsymbol{x},T) = J(\boldsymbol{x}) & \boldsymbol{x}\in\mathbb{R}^n, \end{cases} \tag{7}$$

with Hamiltonian $H\colon [0,T]\times\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ of the form

$$H(t,\boldsymbol{x},\boldsymbol{p}) \doteq \frac{1}{2}\boldsymbol{p}^T C_{pp}(t)\boldsymbol{p} - \frac{1}{2}\boldsymbol{x}^T C_{xx}(t)\boldsymbol{x} - \boldsymbol{p}^T C_{xp}(t)\boldsymbol{x},$$

and terminal condition $J\colon\mathbb{R}^n\to\mathbb{R}$ of the form

$$J(\boldsymbol{x}) \doteq \min_{i\in\{1,\dots,m\}} \left\{ \frac{1}{2}\boldsymbol{x}^T G_i \boldsymbol{x} + \boldsymbol{a}_i^T \boldsymbol{x} + b_i \right\}.$$

The coefficients in H, i.e., $C_{pp}(t)$, $C_{xx}(t)$, and $C_{xp}(t)$, are functions depending on t and taking values in $\mathbb{R}^{n\times n}$. The parameters in J contain matrices $G_i\in S^n$, vectors $\boldsymbol{a}_i\in\mathbb{R}^n$, and scalars $b_i\in\mathbb{R}$, for $i=1,\dots,m$.

Under certain assumptions (see Darbon et al. (2021)), the viscosity solution to (7) can be represented by a function V_{NN} defined by

$$\begin{aligned} V_{NN}(\boldsymbol{x},t) &\doteq \min_{i\in\{1,\dots,m\}} V_i(\boldsymbol{x},t), \\ V_i(\boldsymbol{x},t) &\doteq \frac{1}{2}\boldsymbol{x}^T P_i(t)\boldsymbol{x} + \boldsymbol{q}_i(t)^T\boldsymbol{x} + r_i(t), \end{aligned} \tag{8}$$

where the function $P_i\in C(0,T;S^n)$ solves the following Riccati final value problem (FVP)

$$\begin{cases} \dot{P}_i(t) = P_i(t)^T C_{pp}(t) P_i(t) - P_i(t)^T C_{xp}(t) \\ \qquad\qquad -C_{xp}(t)^T P_i(t) - C_{xx}(t) \quad t\in(0,T), \\ P_i(T) = G_i, \end{cases} \tag{9}$$

the functions $\boldsymbol{q}_i\in C(0,T;\mathbb{R}^n)$ solves the following linear FVP

$$\begin{cases} \dot{\boldsymbol{q}}_i(t) = P_i(t)^T C_{pp}(t)\boldsymbol{q}_i(t) - C_{xp}(t)^T\boldsymbol{q}_i(t) \quad t\in(0,T), \\ \boldsymbol{q}_i(T) = \boldsymbol{a}_i, \end{cases} \tag{10}$$

and the function $r_i\in C(0,T;\mathbb{R})$ solves the following FVP

$$\begin{cases} \dot{r}_i(t) = \dfrac{1}{2}\boldsymbol{q}_i(t)^T C_{pp}(t)\boldsymbol{q}_i(t) \quad t\in(0,T), \\ r_i(T) = b_i. \end{cases} \tag{11}$$

In our implementations, we apply the Runge-Kutta method to solve (9), (10), and (11). The Runge-Kutta method can itself be expressed using a Resnet architecture, so that the implementation for the function V_{NN} defined in (8) can be expressed using the deep neural network illustrated in Fig. 3.

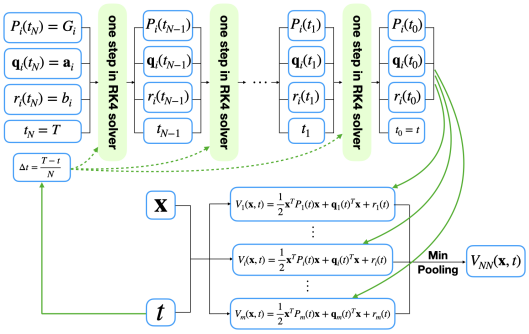


Fig. 3. A deep neural network architecture defined by (8) (where each function V_i is computed using the Runge-Kutta method) that represents the viscosity solution to the HJ PDE (7).

The bottom part of Fig. 3 is a one-layer abstract architecture with a min-pooling activation function. The i-th abstract neuron is given by the function $V_i(\boldsymbol{x},t)$ in (8). The value on the i-th abstract neuron is computed using the top part of Fig. 3, which represents the Runge-Kutta method for solving (9), (10), and (11). Note that the top part of Fig. 3 can be replaced by another numerical method for solving (9), (10), and (11), if the method can be expressed using a neural network architecture.

In Darbon et al. (2021), we provide both theoretical guarantees and numerical results for our proposed deep architecture. For the purpose of illustration here, we extract an example of solving a high-dimensional HJ PDE below. We refer readers to Darbon et al. (2021) for more results on solving HJ PDEs and the corresponding optimal control problems in high dimensions.

We consider the HJ PDE (7) with

$$H(t,\boldsymbol{x},\boldsymbol{p}) \doteq \frac{e^t}{4}\|\boldsymbol{p}\|^2 - \frac{e^{-t}}{4}\|\boldsymbol{x}\|^2 - \frac{1}{2}\langle\boldsymbol{p},\boldsymbol{x}\rangle, \tag{12}$$

and the terminal condition J is defined by

$$J(\boldsymbol{x}) \doteq \min\left\{\Psi_1(\boldsymbol{x}),\Psi_2(\boldsymbol{x}),\Psi_3(\boldsymbol{x}),\Psi_4(\boldsymbol{x})\right\}, \tag{13}$$

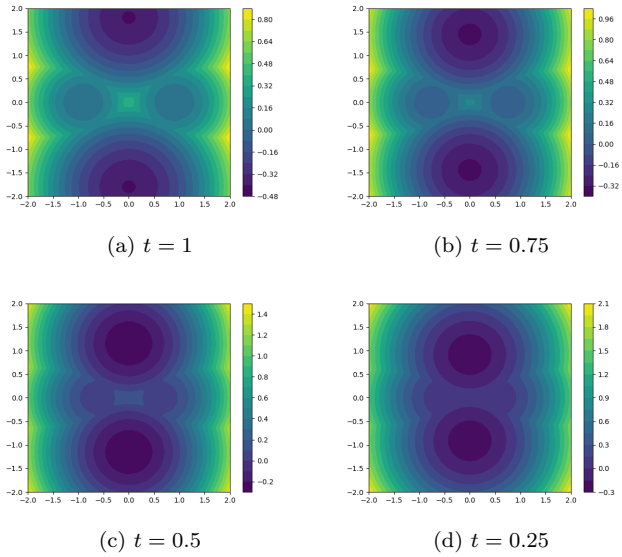
where $\Psi_1,\Psi_2,\Psi_3,\Psi_4\colon\mathbb{R}^n\to\mathbb{R}$ are defined by

$$\begin{aligned} \Psi_1(\boldsymbol{x}) &\doteq 0.5\|\boldsymbol{x}\|^2 + 0.9x_1 + 0.405, \\ \Psi_2(\boldsymbol{x}) &\doteq 0.5\|\boldsymbol{x}\|^2 - 0.9x_1 + 0.405, \\ \Psi_3(\boldsymbol{x}) &\doteq 0.25\|\boldsymbol{x}\|^2 + 0.9x_2 + 0.405, \\ \Psi_4(\boldsymbol{x}) &\doteq 0.25\|\boldsymbol{x}\|^2 - 0.9x_2 + 0.405, \end{aligned}$$

for each $\boldsymbol{x}=(x_1,\dots,x_n)\in\mathbb{R}^n$. We solve a 16-dimensional problem, i.e., we set $n=16$, and the numerical results at earlier times t are shown in Fig. 4. In each figure, we plot two-dimensional slices of the function $\boldsymbol{x}\mapsto V_{NN}(\boldsymbol{x},t)$ for illustration. We consider the points $\boldsymbol{x}=(x_1,x_2,\boldsymbol{0})\in\mathbb{R}^{16}$ where $(x_1,x_2)\in\mathbb{R}^2$ is any grid point in a two-dimensional rectangular domain and $\boldsymbol{0}$ denotes the zero vector in $\mathbb{R}^{14}$. From the numerical results, our proposed deep neural network overcomes the CoD in this example.

REFERENCES

Akian, M., Gaubert, S., and Lakhoua, A. (2008). The max-plus finite element method for solving deterministic optimal control problems: basic properties and convergence analysis. *SIAM Journal on Control and Optimization*, 47(2), 817–848.



(a) $t = 1$ (b) $t = 0.75$

(c) $t = 0.5$ (d) $t = 0.25$

Fig. 4. The viscosity solution V_{NN} to the 16-dimensional HJ PDE (7) with Hamiltonian (12), terminal data (13) and time horizon $T = 1$ is computed using the proposed deep neural network architecture in Fig. 3. The two dimensional slices of V_{NN} at time $t = 1$ (i.e., the terminal cost), $t = 0.75$, $t = 0.5$, and $t = 0.25$ are shown in the subfigures (a), (b), (c), and (d), respectively. The color in each subfigure shows the solution value $V_{NN}(\boldsymbol{x}, t)$, where the spatial variable $\boldsymbol{x}$ is in the form of $(x_1, x_2, \mathbf{0}) \in \mathbb{R}^{16}$ (where $\mathbf{0}$ is the zero vector in $\mathbb{R}^{14}$) for some points $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$ which are represented by x and y axes.

Alla, A., Falcone, M., and Saluzzi, L. (2019). An efficient DP algorithm on a tree-structure for finite horizon optimal control problems. *SIAM Journal on Scientific Computing*, 41(4), A2384–A2406.

Alla, A., Falcone, M., and Volkwein, S. (2017). Error analysis for POD approximations of infinite horizon problems via the dynamic programming approach. *SIAM Journal on Control and Optimization*, 55(5), 3091–3115.

Bachouch, A., Huré, C., Langrené, N., and Pham, H. (2018). Deep neural networks algorithms for stochastic control problems on finite horizon: numerical applications. *arXiv preprint arXiv:1812.05916*.

Bellman, R.E. (1961). *Adaptive control processes: a guided tour*. Princeton university press.

Bertsekas, D.P. (2019). Reinforcement learning and optimal control. *Athena Scientific, Belmont, Massachusetts*.

Bokanowski, O., Garcke, J., Griebel, M., and Klompmaker, I. (2013). An adaptive sparse grid semi-Lagrangian scheme for first order Hamilton-Jacobi Bellman equations. *Journal of Scientific Computing*, 55(3), 575–605.

Darbon, J., Dower, P.M., and Meng, T. (2021). Neural network architectures using min plus algebra for solving certain high dimensional optimal control problems and Hamilton-Jacobi PDEs. *arXiv preprint arXiv:2105.03336*.

Darbon, J., Langlois, G.P., and Meng, T. (2020). Overcoming the curse of dimensionality for some Hamilton-Jacobi partial differential equations via neural network architectures. *Res. Math. Sci.*, 7(3), 20.

Darbon, J. and Osher, S. (2016). Algorithms for overcoming the curse of dimensionality for certain Hamilton-Jacobi equations arising in control theory and elsewhere. *Res Math Sci Research in the Mathematical Sciences*, 3(19), 1–26.

Darbon, J. and Meng, T. (2021). On some neural network architectures that can represent viscosity solutions of certain high dimensional Hamilton–Jacobi partial differential equations. *Journal of Computational Physics*, 425, 109907.

Dolgov, S., Kalise, D., and Kunisch, K. (2019). A tensor decomposition approach for high-dimensional Hamilton-Jacobi-Bellman equations. *arXiv preprint arXiv:1908.01533*.

Fleming, W. and McEneaney, W. (2000). A max-plus-based algorithm for a Hamilton–Jacobi–Bellman equation of nonlinear filtering. *SIAM Journal on Control and Optimization*, 38(3), 683–710.

Hiriart-Urruty, J.B. and Lemaréchal, C. (1993a). *Convex Analysis and Minimization Algorithms I: Fundamentals*, volume 305. Springer-Verlag Berlin Heidelberg.

Hiriart-Urruty, J.B. and Lemaréchal, C. (1993b). *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, volume 306. Springer science & business media.

Horowitz, M.B., Damle, A., and Burdick, J.W. (2014). Linear Hamilton Jacobi Bellman equations in high dimensions. In *53rd IEEE Conference on Decision and Control*, 5880–5887. IEEE.

Jiang, F., Chou, G., Chen, M., and Tomlin, C.J. (2016). Using neural networks to compute approximate and guaranteed feasible Hamilton-Jacobi-Bellman PDE solutions. *arXiv preprint arXiv:1611.03158*.

Jin, P., Zhang, Z., Zhu, A., Tang, Y., and Karniadakis, G.E. (2020). SympNets: Intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems. *Neural Networks*, 132, 166–179.

Kalise, D., Kundu, S., and Kunisch, K. (2019). Robust feedback control of nonlinear PDEs by numerical approximation of high-dimensional Hamilton-Jacobi-Isaacs equations. *arXiv preprint arXiv:1905.06276*.

Kang, W. and Wilcox, L.C. (2017). Mitigating the curse of dimensionality: sparse grid characteristics method for optimal feedback control and HJB equations. *Computational Optimization and Applications*, 68(2), 289–315.

Kunisch, K., Volkwein, S., and Xie, L. (2004). HJB-POD-based feedback design for the optimal control of evolution problems. *SIAM Journal on Applied Dynamical Systems*, 3(4), 701–722.

McEneaney, W. (2006). *Max-plus methods for nonlinear control and estimation*. Springer Science & Business Media.

Nakamura-Zimmerer, T., Gong, Q., and Kang, W. (2019). Adaptive deep learning for high-dimensional Hamilton-Jacobi-Bellman equations. *arXiv preprint arXiv:1907.05317*.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28), 11478–11483.

Yegorov, I. and Dower, P.M. (2017). Perspectives on characteristics based curse-of-dimensionality-free numerical approaches for solving Hamilton–Jacobi equations. *Applied Mathematics & Optimization*, 1–49.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Commutative and Non-Commutative Polynomial Optimization in Quantum Optimal Control (Abstract)

Jakub Marecek * **Jiri Vala** **

* *Czech Technical University, Prague, the Czech Republic.*
** *Maynooth University, Maynooth, Ireland.*

Keywords: Optimal control theory, control of quantum and Schroedinger systems, Sum-of-squares, convex optimization, industrial applications of optimal control

1. INTRODUCTION

Quantum optimal control has numerous applications, within quantum computing, laser control of chemical reactions, and nuclear magnetic resonance. In quantum computing, better quantum optimal control provides faster and more accurate two-qubit gates, and multi-level operations in general, eventually allowing for fault-tolerant quantum computation.

There are many excellent results, which establish conditions for controllability, as surveyed in D'Alessandro (2007) and Jurdjevic (2016), but constructive, algorithmic approaches still leave space for improvement. For the control of systems involving more than three levels, including the pulse-shaping for two-qubit gates, there are no deterministic, globally convergent solvers. There are two key challenges. First, most formulations seem to assume commutativity of the Hamiltonian at different times. Second, the corresponding quantum optimal control on an N-level system is non-convex for $N \geq 4$, but only heuristics based on first-order optimality conditions are employed.

We have recently addressed the issues of non-commutativity and non-convexity of the problem by employing Magnus expansion (Magnus, 1954; Blanes et al., 2009) and tools from non-commutative polynomial optimisation (Pironio et al., 2010; Burgdorf et al., 2016). In addressing the non-commutativity of the problem, this improves most directly upon the work of Schutjens et al. (2013) and Theis et al. (2016), who consider the lowest-order term of the Magnus expansion, also known as the average Hamiltonian, and derive conditions for all other terms being zero. In contrast to their approach, we consider an arbitrary number of terms in the Magnus expansion. Our work complements research on Magnus expansion in numerical integration of the Schrödinger equation (Blanes et al., 2009; Singh, 2018; Kopylov, 2019, e.g.), which however have not been developed in the context of quantum control so far. In addressing the non-convexity of the problem, we utilise a hierarchy of progressively stronger convexifications. This improves upon essentially all related work on quantum optimal control, which guarantees only monotonic convergence to first-order critical points or local minima of a non-convex optimisation problem based on Pontryagin's maximum principle.

2. THE PROBLEM

Let us consider a finite N-dimensional quantum system whose time-evolution is governed by a Schrödinger equation. Given an initial condition $\hat{U}(0) = \hat{I}$, where $\hat{I}$ is a unit matrix in $\mathbb{C}^{N\times N}$, a terminal time $T > 0$, and a target unitary $\hat{U}^* \in U(N) \subset \mathbb{C}^{N\times N}$, where $U(N)$ is the Lie group of $N \times N$ unitary operators or matrices, we aim to control a time-dependent Hamiltonian $\hat{H}(t)$ over time $t \in [0, T]$. That is, we seek a particular solution of the initial value problem for the Schrödinger equation:

$$\frac{\partial}{\partial t}\hat{U}(t) = \hat{A}(t)\hat{U}(t) \tag{1}$$

where $\hat{A}(t) = \hat{H}(t)/i\hbar$ can explicitly be written in terms of controls $u_j(t) : [0, T] \to \mathbb{R}$ as

$$\hat{A}(t) = \sum_j u_j(t)\ \hat{H}_j/i\hbar. \tag{2}$$

In particular, we seek a solution that is optimal with respect to a given functional J, while using controls $\{u_j(t)\}$ constrained to some set Υ. Formally, the quantum optimal control problem reads:

$$\min_{\hat{U}(t),\{u_j(t)\}\in\Upsilon} J\left(\hat{U}(t), \{u_j(t)\}\right) \tag{3}$$
$$\text{s.t. } \frac{\partial}{\partial t}\hat{U}(t) = \left[\sum_j u_j(t)\ \hat{H}_j/i\hbar\right]\hat{U}(t),$$
$$\hat{U}(0) = \hat{I}.$$

where J is the (objective) functional for the control problem, which is polynomially or semidefinite representable (Helton and Vinnikov, 2007), and Υ is a polynomially representable set.

3. OUR APPROACH

It is well known that initial value problem (1) has a solution in the form of the Magnus expansion (Magnus, 1954; Blanes et al., 2009):

$$\Omega(T) = \sum_{m=1}^{\infty} \Omega_m(T), \tag{4}$$

where the individual terms in the series require evaluations of increasingly more complex integrals involving nested commutators. When the series $\Omega_m(T)$ is absolutely convergent, then $\hat{U}(t)$ can be written in the form

$$\hat{U}(t) = \exp \Omega(t). \tag{5}$$

In a key insight of this paper, we show that the nested commutators between the Hamiltonian at different times are instrumental in extending the reachable set. This expansion is accomplished via two distinct mechanisms. First, the nested commutators generate new linearly independent elements of the Lie algebra $\mathfrak{su}(N)$ and hence increase the dimension of the reachable set. The operator controllability is accomplished when this dimension reaches $N^2 - 1$.

The second mechanism is related to the controls $\{u_j(t)\}$, which figure in coefficients of the Lie algebra elements that are generated by the k terms of the Magnus expansion. The controls are, in general, arbitrary functions of time which allows for an arbitrary linear combination of the Lie algebra elements. Hence the two mechanisms result in both expanding the dimension of the reachable set as a Lie algebra and also in its dense cover by control functions and their integrals. We denote the expanded reachable set obtained with the lowest m terms in the Magnus series by $\mathrm{ME}(R(\hat{U}(0)), m)$.

Specifically, in the case of a time-dependent Hamiltonian, one need not consider only generators of the Lie group as they appear in $\hat{H}(t)$, but one can also consider the commuting relations obtained by Magnus expansion: (i) A necessary condition for reachability of any $\hat{U}(T) \in SU(N)$ from $\hat{U}(0) = \hat{I}$ to $\hat{U}(T)$, considering Magnus expansion is that the dimension of the Lie algebra generated by Magnus expansion has dimension $N^2 - 1$. (ii) A necessary and sufficient condition for the existence of time T^* such that for all $T > T^*$ one has exact-time operator controllability is that the dimension of the Lie algebra generated by Magnus expansion has dimension $N^2 - 1$.

It is to be pointed out that the validity of our approach is limited by the convergence of the Magnus expansion. This has been studied extensively, with the most recent result provided by Moan and Niesen (2008).

4. THE MAIN RESULT

Our second insight is that this approach can be made constructive, considering that for any number m of terms in the Magnus expansion, one obtains a non-commutative polynomial optimisation problem (NCPOP), which can be solved by solving a sequence (Pironio et al., 2010, cf.) of natural linear matrix inequalities (Boyd et al., 1994) in the original variables and additional variables for non-linear monomials, based on the Sums of Squares theorem of Helton (2002) and McCullough (2001).

Hence, under mild assumptions, for any initial state $\hat{U}(0)$, for any lower bound $\underline{m}$ on the number of terms in the Magnus expansion, for any target state in the expanded reachable set $\mathrm{ME}(R(\hat{U}(0)), \underline{m})$, and any error $\epsilon > 0$, there is a number of terms $m(\epsilon) \geq \underline{m}$ such that ϵ-optimal control with respect to any polynomially-representable functional can be extracted from the solution of a certain convex optimisation problem in the model of Blum et al. (1989).

Notice that we use Magnus expansion in two ways here: First, $\underline{m}$ steps in the Magnus expansion guarantee we can reach any target in $\mathrm{ME}(R(\hat{U}(0))$. Second, we need $m(\epsilon) \geq \underline{m}$ number of steps to achieve the convergence within ϵ error introduced by the Magnus expansion. We also utilise recent results on the robustness of the GNS construction to small errors. Indeed, we can apply (Klep et al., 2018, Theorem 3.2) directly, if there are no constraints Υ.

Now notice that one could also consider an unconstrained problem, where $\sum_{m=1}^{\overline{m}} \Omega_m(T)$ is compared to $\log U^*$ directly, measuring a distance between the target element of the Lie algebra given by $\log U^*$ and the element of the Lie algebra emerging from the Magnus expansion $\tilde{\Omega} = \sum_{m=1}^{\overline{m}} \Omega_m(T)$ of the control problem. There, we can rewrite the expression for $\Omega_m(T)$ as $\sum_i \hat{O}_i \tilde{F}_i(\{\tilde{u}_j\}(t_1, \ldots, t_{m+1}))$. The operators $\hat{O}_i$ result from the commutators of the Magnus expansion which involve the Hamiltonians $\hat{H}_j$ at different times. The function $\tilde{F}_i(\tilde{u}_j\}(t_1, \ldots, t_{m+1}))$ is a polynomial of the time-dependent controls $u_j(t)$ at different times $t_1, \ldots, t_{m+1}$, originating from the same commutator, and after an appropriate discretization (sub-sampling) of time $\{u_j\} \to \{\tilde{u}_j\}$. The functions $\{\tilde{F}_i\}$ represent coefficients in the Lie algebra associated with the operator $\{O_i\}$ which, if the system is controllable, constitute the complete set of generators of the Lie algebra. In this formulation, the control problem reduces to optimization of these coefficients, and in particular minimization of the difference between these coefficients and those of $\log U^*$. Since $\log U^*$ is not unique, one needs to consider this comparison modulo 2π which in turn may weaken the guarantees of finding the global minimum. At the same time, however, the problem becomes a commutative polynomial optimization problem (POP).

5. CONCLUSIONS

We have presented an approach to quantum optimal control that exhibits global convergence, in theory, and relies on non-trivial but well-developed tools from non-commutative geometry and mathematical optimisation, in practice. In contrast to other quantum control approaches, the use of Magnus expansion provides the proper solution of the initial value problem of the Schrödinger equation involving time-dependent Hamiltonian. This has a significant impact on the controllability of quantum systems in that it expands the reachable set both in its dimension and volume. This opens new avenues for research and engineering in quantum control and its applications such as quantum computing.

Quantum optimal control can significantly enhance quantum computing in the context of Noisy Intermediate Scale Quantum (NISQ) computing devices. Immediately, one can improve fidelity of two-qubit gates. In principle, one can also replace the application of an entire quantum circuit with a control signal. This would make it possible to move beyond the quantum circuit model and the associated intricacy of approximate compiling and swap mapping to accommodate connectivity constraints present in many qubit technologies.

ACKNOWLEDGEMENTS

We would like to thank Daniel Egger, Didier Henrion, Christiane Koch, and Vyacheslav Kungurtsev for their insightful comments.

REFERENCES

Blanes, S., Casas, F., Oteo, J.A., and Ros, J. (2009). The magnus expansion and some of its applications. *Physics Reports*, 470, 151 – 238.

Blum, L., Shub, M., and Smale, S. (1989). On a theory of computation and complexity over the real numbers: *np*- completeness, recursive functions and universal machines. *Bull. Amer. Math. Soc. (N.S.)*, 21(1), 1–46. URL `https://projecteuclid.org:443/euclid.bams/1183555121`.

Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*, volume 15. Siam.

Burgdorf, S., Klep, I., and Povh, J. (2016). *Optimization of polynomials in non-commuting variables*. Springer.

D'Alessandro, D. (2007). *Introduction to quantum control and dynamics*. Chapman and Hall/CRC.

Helton, J.W. (2002). "positive" noncommutative polynomials are sums of squares. *Annals of Mathematics*, 156(2), 675–694.

Helton, J.W. and Vinnikov, V. (2007). Linear matrix inequality representation of sets. *Communications on Pure and Applied Mathematics*, 60(5), 654–674.

Jurdjevic, V. (2016). *Optimal control and geometry: integrable systems*, volume 154. Cambridge University Press.

Klep, I., Povh, J., and Volcic, J. (2018). Minimizer extraction in polynomial optimization is robust. *SIAM Journal on Optimization*, 28(4), 3177–3207.

Kopylov, N. (2019). *Magnus-based geometric integrators for dynamical systems with time-dependent potentials*. Ph.D. thesis, Universitat Politecnica de Valencia.

Magnus, W. (1954). *Comm. Pure Appl. Math.*, 7, 649.

McCullough, S. (2001). Factorization of operator-valued polynomials in several non-commuting variables. *Linear Algebra and its Applications*, 326(1-3), 193–203.

Moan, P.C. and Niesen, J. (2008). Convergence of the magnus series. *Foundations of Computational Mathematics*, 8(3), 291–301.

Pironio, S., Navascués, M., and Acín, A. (2010). Convergent relaxations of polynomial optimization problems with noncommuting variables. *SIAM Journal on Optimization*, 20(5), 2157–2180. doi:10.1137/090760155.

Schutjens, R., Dagga, F.A., Egger, D.J., and Wilhelm, F.K. (2013). Single-qubit gates in frequency-crowded transmon systems. *Phys. Rev. A*, 88, 052330. doi:10.1103/PhysRevA.88.052330. URL `https://link.aps.org/doi/10.1103/PhysRevA.88.052330`.

Singh, P. (2018). *High accuracy computational methods for the semiclassical Schrödinger equation*. Ph.D. thesis, University of Cambridge. doi:10.17863/CAM.22064.

Theis, L.S., Motzoi, F., and Wilhelm, F.K. (2016). Simultaneous gates in frequency-crowded multilevel systems using fast, robust, analytic control shapes. *Phys. Rev. A*, 93, 012324. doi:10.1103/PhysRevA.93.012324. URL `https://link.aps.org/doi/10.1103/PhysRevA.93.012324`.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Sparse System Identification with Kernel Regularization ⋆

Masaaki Nagahara * **Yusuke Fujimoto** ** **Yutaka Yamamoto** ***

*

**

***

Abstract: In this article, we propose a novel system identification method for stable and sparse linear time-invariant systems. We adopt kernel-based regularization to take a priori information, such as the decay rate, of the target system into account. For promoting sparsity, we introduce the minimax concave penalty function, which is known to promote sparser results than the standard ℓ^1 penalty. The estimation problem is shown to be reduced to a convex optimization problem, which can be efficiently solved by the forward-backward algorithm. We show a numerical example of delayed FIR (finite impulse response) system identification to illustrate the effectiveness of the proposed method.

System identification, kernel method, sparse optimization, convex optimization, minimax concave penalty.

1. INTRODUCTION

The accuracy of system identification depends highly on how a priori information is utilized. The k

is one of the most effective methods to take such information into account, as discussed in Pillonetto et al. (2014). For example, in Pillonetto and De Nicolao (2010); Chen et al. (2012), it was shown that the kernel method can significantly improve the accuracy of system identification of a stable linear system by adopting the exponential convergence of impulse response. More recently, the method has been extended to a priori information on the DC gain (Fujimoto and Sugie, 2018), the frequency domain decay characteristics (Fujimoto, 2021b,a), and the relative degree (Fujimoto et al., 2017).

On the other hand, the notion of of vectors and functions also plays an important role in control systems design, e.g, maximum hands-off control (Nagahara et al., 2016; Nagahara, 2020, 2021), and sparse system identification (Chen et al., 2009; Fattahi and Sojoudi, 2018). In this article, we study identification of sparse impulse response that has only a few nonzero coefficients. For this, we introduce sparsity-promoting penalty function to the kernel regularization to take account of sparsity and other system properties at the same time. In particular, we propose to use the minimax concave penalty function discussed in Selesnick (2017) as a sparsity prior, instead of the standard ℓ^1 penalty used in Chen et al. (2009); Fattahi and Sojoudi (2018). Although the minimax concave penalty function is not convex, it promotes sparsity more than the ℓ^1 penalty. Moreover, by choosing appropriate hyper parameters, the cost function to be minimized becomes convex, and the minimization problem can be solved by an efficient algorithm. By a numerical example, we show the effectiveness of the proposed method in particular for an FIR (finite-impulse-response) system with a large delay.

Note that the approach is different from the work of Chen et al. (2014) that proposes to sparsify the hyper-parameter vector in multiple kernels by maximizing the marginal likelihood, which is not a convex optimization, and just a local optimal solution is obtained.

For vector $v = [v_1, \dots, v_m]^\top$, $\|v\|_1$ is the ℓ^1 norm defined by $\|v\|_1 \triangleq |v_1| + \cdots + |v_m|$, and $\|v\|_2$ is the ℓ^2 norm defined by $\triangleq \sqrt{v^\top v}$, where $\top$ denotes the transpose. A matrix A whose (i,j)-th element is A_{ij} is described as $A = [A_{ij}]$. For a square matrix A, $\|A\|$ denotes the maximum singular value of A. For symmetric matrix A, we write $A \geq 0$ when A is positive semi-definite.

2. PROBLEM FORMULATION

In this paper, we consider the FIR (finite impulse response) model described by

$$y_k = \sum_{i=0}^{m-1} g_i u_{k-i} + \epsilon_k, \quad k = 0, 1, 2, \dots, \tag{1}$$

where $\{u_k\}$ and $\{y_k\}$ are respectively the input and output sequences, $\{\epsilon_k\}$ is noise, and $\{g_i\}$ is the impulse response. The problem is to identify the impulse response $\{g_i : i =$

⋆ This work was partly supported by JSPS KAKENHI Grant Numbers JP20H02172 and JP20K21008.

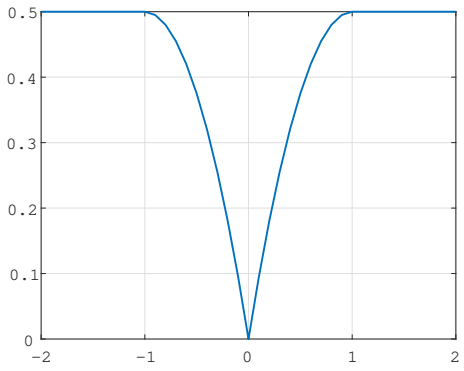


Fig. 1. Minimax concave penalty function ψ_B with $B = 1$.

$0, 1, \ldots, m-1\}$ from the input/output data $\{(u_k, y_k) : k = 0, 1, \ldots, N\}$. For this, we define the following vectors and matrix:

$$y \triangleq \begin{bmatrix} y_0 \\ \vdots \\ y_N \end{bmatrix}, \ \epsilon \triangleq \begin{bmatrix} \epsilon_0 \\ \vdots \\ \epsilon_N \end{bmatrix}, \ g \triangleq \begin{bmatrix} g_0 \\ \vdots \\ g_{m-1} \end{bmatrix},$$

$$U \triangleq \begin{bmatrix} u_0 & 0 & \ldots & 0 \\ u_1 & u_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ u_{N-1} & u_{N-2} & \ldots & u_{N-m} \end{bmatrix}.$$

Then, we consider the squared ℓ^2 error $E(g)$ as

$$E(g) \triangleq \frac{1}{2}\|y - Ug\|_2^2. \tag{2}$$

The minimizer of $E(g)$ is the least-square solution, which is used when we have no prior information on g. However, it may cause overfitting (Bishop, 2006) in particular when the parameter size m is large. To avoid this, we adopt with the following regularization term:

$$\Omega(g) \triangleq \frac{\mu}{2} g^\top K^{-1} g + (1-\mu)\psi_B(g), \tag{3}$$

where $K = [K_{ij}]$ is a positive definite kernel matrix, and ψ_B is the (generalized) minimax concave penalty function (Selesnick, 2017) defined by

$$\psi_B(g) \triangleq \|g\|_1 - \min_h \left\{ \|h\|_1 + \frac{1}{2}\|B(g-h)\|_2^2 \right\}, \tag{4}$$

where B is a matrix, which is a hyper-parameter, such that $B^\top B$ is non-singular. Figure 2 shows the curve of the 1-dimensional minimax concave penalty function ψ_B with $B = 1$.

The kernel matrix is introduced to take a priori knowledge on the target system into account. For the simplicity of discussion, m is assumed to be even in the rest of this section. This paper uses the High-Frequency Decay (HFD) kernel (Fujimoto, 2021b) which is given by

$$K = F^\top \begin{bmatrix} K^{\mathrm{re}} & 0 \\ 0 & K^{\mathrm{im}} \end{bmatrix} F \in \mathbb{R}^{m \times m}, \tag{5}$$

where $K^{\mathrm{re}} = [K_{ij}^{\mathrm{re}}] \in \mathbb{R}^{\left(\frac{m}{2}+1\right)\times\left(\frac{m}{2}+1\right)}$ and $K^{\mathrm{im}} = [K_{ij}^{\mathrm{im}}] \in \mathbb{R}^{\left(\frac{m}{2}-1\right)\times\left(\frac{m}{2}-1\right)}$ are respectively given by

$$K_{ij}^{\mathrm{re}} = k_{\mathrm{HFD}}(\omega_{i-1}, \omega_{j-1}), \ K_{ij}^{\mathrm{im}} = k_{\mathrm{HFD}}(\omega_i, \omega_j), \tag{6}$$

$$k_{\mathrm{HFD}}(\omega_i, \omega_j) = \eta_1 \min\left\{ \frac{1}{\left(\omega_i^2 + \eta_2\right)^d}, \frac{1}{\left(\omega_j^2 + \eta_2\right)^d} \right\}, \tag{7}$$

$$\omega_i = \frac{2\pi}{N} i. \tag{8}$$

Also $F \in \mathbb{R}^{m\times m}$ is defined as

$$F = \begin{bmatrix} F^{\mathrm{re}} \\ F^{\mathrm{im}} \end{bmatrix}, \tag{9}$$

where $F^{\mathrm{re}} = [F_{ij}^{\mathrm{re}}] \in \mathbb{R}^{\left(\frac{m}{2}+1\right)\times m}$ and $F^{\mathrm{im}} = [F_{ij}^{\mathrm{im}}] \in \mathbb{R}^{\left(\frac{m}{2}-1\right)\times m}$ are given by

$$F_{ij}^{\mathrm{re}} = \cos\left(\frac{2\pi(i-1)(j-1)}{m}\right), F_{ij}^{\mathrm{im}} = -\sin\left(\frac{2\pi i(j-1)}{m}\right). \tag{10}$$

It is known that the estimated impulse response with the HFD kernel shows high-frequency decay property (Fujimoto, 2021b).

The minimax concave penalty function (4) promotes sparsity of the estimated impulse response. Although this is non-convex (see Figure 2), the cost function

$$J(g) \triangleq E(g) + \lambda\Omega(g), \tag{11}$$

with $\lambda > 0$ becomes convex if we appropriately choose B. In fact, we have the following proposition.

Assume

$$U^\top U + \lambda\left(\mu K^{-1} - (1-\mu)B^\top B\right) \geq 0. \tag{12}$$

Then $J(g)$ is convex over $\mathbb{R}^m$.

Proof: Define

$$\tilde{U} \triangleq \begin{bmatrix} U \\ V \end{bmatrix}, \quad \tilde{y} \triangleq \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad \tilde{\lambda} \triangleq \lambda(1-\mu), \tag{13}$$

where V is a matrix satisfying $V^\top V = \lambda\mu K^{-1}$. Then it is easily shown that

$$J(g) = \frac{1}{2}\|\tilde{y} - \tilde{U}g\|_2^2 + \tilde{\lambda}\psi_B(g). \tag{14}$$

Then from Theorem 1 of Selesnick (2017), $J(g)$ is convex if $\tilde{\lambda}B^\top B \leq \tilde{U}^\top\tilde{U}$, which is equivalent to (12). □

By combining minmax concave penalty function and the HFD kernel, the estimated impulse response is expected to have sparsity in time domain, and high-frequency decay property in frequency domain.

3. OPTIMIZATION ALGORITHM

To satisfy (12), we choose B such that $\tilde{\lambda}B^\top B = \gamma\tilde{U}^\top\tilde{U}$ with $\gamma \in (0, 1)$. Then, minimizing $J(g)$ is equivalent to the following saddle point problem:

$$\min_g \max_h \frac{1}{2}\|\tilde{y} - \tilde{U}g\|_2^2 + \tilde{\lambda}\|g\|_1 - \tilde{\lambda}\|h\|_1 - \frac{\gamma}{2}\|\tilde{U}(g-h)\|_2^2. \tag{15}$$

This is efficiently solved by the forward-backward algorithm (see Selesnick (2017) and Bauschke and Combettes (2011) for details):

$$\begin{aligned} w[k] &= g[k] - c\tilde{U}^\top\left\{\tilde{U}\left(g[k] + \gamma(h[k] - g[k])\right) - \tilde{y}\right\}, \\ u[k] &= h[k] - c\gamma\tilde{U}^\top\tilde{U}(h[k] - g[k]), \\ g[k+1] &= S_{c\tilde{\lambda}}(w[k]), \\ h[k+1] &= S_{c\tilde{\lambda}}(u[k]), \quad k = 0, 1, 2, \ldots, \end{aligned} \tag{16}$$

where c is a step size of this iteration that satisfies $0 < c < \max\{1, \gamma/(1-\gamma)\}\|\tilde{U}^\top\tilde{U}\|$, and S_a is the soft-thresholding function with parameter $a > 0$ defined by

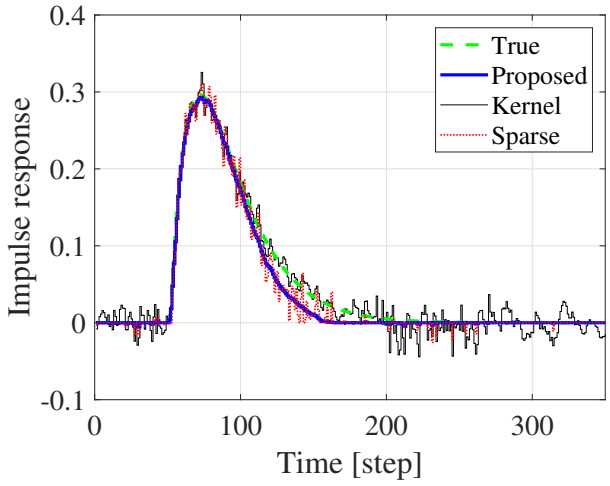


Fig. 2. Impulse response: true (broken line), estimated by the HFD kernel (thin-solid line), sparse regularization (dotted line), and by the proposed method (thick-solid line)

$[S_a(v)]_i \triangleq \mathrm{sign}(v)\max\{|v_i| - a, 0\}$, where $[S_a(v)]_i$ and v_i are respectively the i-th elements of $S_a(v)$ and v. We note that this function is the proximal operator of the ℓ^1 norm; see Nagahara (2020) for details.

4. NUMERICAL EXAMPLE

Let us consider the following continuous-time transfer function:

$$P(s) = \frac{100}{(s+3)(s+2)}e^{-s}. \tag{17}$$

Then, we discretize this by the zero-order hold discretization to obtain a discrete-time system $P_d(z)$. The goal is to estimate the impulse response of $P_d(z)$. We should note that the impulse response is exactly zero in $[0, 1]$ due to the delay e^{-s}. In this example, we assume we know the impulse response is delayed by some delay time, which is not known, and hence it can be assumed to be sparse in the time domain.

For this, we set the length of estimated impulse response to be $m = 350$. We set the input signal u_k in (1) taking ± 1 independently drawn from the Bernoulli distribution with equal probability 0.5. We also add noise ϵ_k in (1) independently drawn from the normal distribution with mean 0 and variance 0.1. The hyperparameters are tuned to $\eta_1 = 10, \eta_2 = 2, \lambda = 50$ and $\mu = 0.8$.

Figure 4 shows the results with the proposed, HFD kernel ($\mu = 1, \lambda = 50$), and sparse regularization($\mu = 0, \lambda = 10$). The proposed method estimate the delay well by using the sparsity of impulse response, and at the same time make the impulse response smooth by using the high-frequency decay property. On the other hand, the sparse regularization (dotted line) does not gives smooth impulse response, and the HFD kernel regularization (thin-solid line) can not estimate the delay of the system. This result indicates that using both the sparsity in time-domain and high-frequency decay in frequency domain can improve the identification accuracy for delayed systems in particular.

5. CONCLUSION

In this paper, we have proposed a novel method of sparse system identification with kernel regularization. We have adopted the minimax concave penalty function for sparsity promoting regularization. By numerical example, the proposed method has an advantage of estimating delayed systems.

REFERENCES

Bauschke, H.H. and Combettes, P.L. (2011). *x* . Springer.

Bishop, C.M. (2006). . Springer.

Chen, T., Ohlsson, H., and Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes–Revisited. , 48(8), 1525–1535.

Chen, T., Andersen, M.S., Ljung, L., Chiuso, A., and Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. , 59(11), 2933–2945.

Chen, Y., Gu, Y., and Hero, A.O. (2009). Sparse lms for system identification. In , 3125–3128.

Fattahi, S. and Sojoudi, S. (2018). Data-driven sparse system identification. In , 462–469.

Fujimoto, Y. (2021a). Kernel regularization for low-frequency decay systems. In , 308–3023.

Fujimoto, Y. (2021b). Kernel regularization in frequency domain: Encoding high-frequency decay property. , 5, 367–372.

Fujimoto, Y., Maruta, I., and Sugie, T. (2017). Extension of first-order stable spline kernel. In *W* , 15481–15486.

Fujimoto, Y. and Sugie, T. (2018). Kernel-based impulse response estimation with a priori knowledge on the DC gain. , 2(4), 713–718.

Nagahara, M. (2020). . Now Publishers.

Nagahara, M., Quevedo, D.E., and Nešić, D. (2016). Maximum hands-off control: a paradigm of control effort minimization. , 61(3), 735–747.

Nagahara, M. (2021). Sparse control for continuous-time systems. .

Pillonetto, G. and De Nicolao, G. (2010). A new kernel-based approach for linear system identification. , 46(1), 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. , 50(3), 657–682.

Selesnick, I. (2017). Sparse regularization via convex analysis. , 65(17), 4481–4494.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Noncommutative Real Algebraic Geometry and Quantum Games

J. William Helton *

* *UC San Diego, La Jolla Ca 92093 USA , (e-mail: helton@math.ucsd.edu)*

Abstract: The last two decades produced a substantial noncommutative (in the free algebra) real and complex algebraic geometry. The aim of the subject is to develop a systematic theory of equations and inequalities for noncommutative polynomials in operator variables. A problem leading very directly to such equations and inequalities is finding good quantum strategies for games. The talk will focus on quantum games and present recent results done jointly with Adam Bene Watts, Igor Klep, Vern Paulsen Mousavi, Nezhadi, Russel, and Zehong Zhao.

Keywords: Quantum Games, Noncommutative Real Algebraic Geometry, Noncommutative optimization, SAT, Operator Theory Techniques

1. EXTENDED ABSTRACT

The talk will select topics from quantum games papers posted on arXiv in the last 2 years by the speaker as well as papers in progress.

A k-player game called kXOR can be played with a classical strategy or with a quantum strategy (if you have tons of equipment). 2XOR arose as a generalization of the famous Bell inequalities which proved quantum entanglement existed. The game is cooperative in that the k players are trying to improve their joint score and 1 is the max score any game can ever achieve; hence a game for which a score of 1 is possible is called a Cperfect or Qperfect game depending on whether the perfect strategy is Quantum or Classical. 2XOR was well understood by Tsielrson in the 1980's. Work with Watts focuses on 3XOR: we give a polynomial time algorithm for either producing a Qperfect strategy or alerting that none exists. Such existence was previously not known to be decidable.

Another line of work with Klep and Bene Watts shows how to associate (about) any quantum game with a left *-ideal in the algebra $\mathcal{P}$ of all noncommutative polynomials in the appropriate number of variables. Earlier Vern Paulsen with a variety of collaborators showed that any synchronous game G can be associated to a C^* algebra A, which amounts to associating G to a 2 sided ideal I with A being the quotient of $\mathcal{P}$ by I. This is extended to general games by our association of a left ideal to any game.

Given a game, each strategy employed has a score and the main problem is to maximize this score. A quantum strategy consists of operators (matrices) on a Hilbert space (typically n dimensional). So typically we fix n and compute a local optium $\hat{X}$, of course we wish we knew if this was a global optimium, but an even more primative question is: does a higher max exist at some dimension $\tilde{n} > n$? Helton, Paulsen Mousavi, Nezhadi, Russel, give a necessary condition for the answer to be no.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Energy Conversion and Dissipativity

Extended Abstract

Arjan van der Schaft * **Dimitri Jeltsema** **

* *Bernoulli Institute for Mathematics, Computer Science and AI, Jan C. Willems Center for Systems and Control, University of Groningen, the Netherlands (e-mail: a.j.van.der.schaft@rug.nl)*
** *Systems and Control Engineering & Reliable Power Supply, HAN University of Applied Sciences, Arnhem, the Netherlands (e-mail: d.jeltsema@han.nl)*

Abstract: The Second Law of thermodynamics implies that no thermodynamic system with a single heat source at constant temperature can convert heat into mechanical work in a repeatable manner. First, we note that this is equivalent to cyclo-passivity at the mechanical port of the thermodynamic system which is constrained by constant temperature at the thermal port. Second, we address the general system-theoretic question which physical systems with two power ports share this property, called one-port cyclo-passivity. Recently, sufficient conditions for one-port cyclo-passivity have been obtained, based on the structure of the interconnection matrix in the port-Hamiltonian formulation. We elaborate on these conditions and provide some extensions. Next we focus on control strategies which go beyond the classical Carnot cycle in order to convert energy in case of one-port cyclo-passivity, and apply this to a number of multiphysics systems.

Keywords: Energy control, passivity, thermodynamics, multiphysics systems
AMS Subject Classification: 93A10, 80A10

1. INTRODUCTION

Energy conversion is a common phenomenon in many multiphysics systems: electro-mechanical, electro-chemical, electro-kinetic, thermal-mechanical, thermal-chemical, thermal diffusion, etc. (see e.g. Kondepudi, Prigogine (2015)). On the other hand, the Second Law of thermodynamics states that *heat* cannot be freely converted into *mechanical work*. Thus energy at the thermal port can*not* be freely converted into energy at the mechanical port. Recently it has been shown in Van der Schaft, Jeltsema (2022, 2021) that this same phenomenon also occurs in quite a few other multiphysics systems. Furthermore, it has been shown that the presence of this phenomenon is closely related to the structure of the interconnection matrix in the port-Hamiltonian formulation of the system.

In fact, consider a (multi-)physical system with two power ports (u_1, y_1) and (u_2, y_2), as schematically depicted in Figure 1.

Fig. 1. Two-port physical system.

We assume throughout that the system is *cyclo-passive*; i.e., there is no internal energy creation and thus

$$\oint \left(y_1^\top(t)u_1(t) + y_2^\top(t)u_2(t)\right) dt \geq 0 \tag{1}$$

for all trajectories bringing the state back to its original value; see e.g. Willems (1972); Van der Schaft (2017, 2020). Now suppose that the state vector x in the port-Hamiltonian formulation of the system, see e.g. Van der Schaft, Jeltsema (2014); Van der Schaft (2017), can be split as

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

in such a way that the port-Hamiltonian equations take the form

$$\begin{aligned} \dot{x}_1 &= J_1(x_1, x_2)e_1 - \mathcal{R}_1(x_1, x_2, e_1) + u_1, \\ \dot{x}_2 &= J_2(x_1, x_2)e_2 - \mathcal{R}_2(x_1, x_2, e_1) + G_2(x_1, x_2)u_2, \end{aligned} \tag{2}$$

together with the corresponding output equations

$$\begin{aligned} y_1 &= e_1, \quad e_1 = \frac{\partial H}{\partial x_1}(x_1, x_2), \\ y_2 &= G_2^\top(x_1, x_2)e_2, \quad e_2 = \frac{\partial H}{\partial x_2}(x_1, x_2). \end{aligned} \tag{3}$$

Here H is the Hamiltonian (total stored energy) of the system, $J_1(x_1, x_2)$ and $J_2(x_1, x_2)$ are skew-symmetric matrices, and the mappings $\mathcal{R}_1(x_1, x_2, e_1), \mathcal{R}_2(x_1, x_2, e_1)$, modeling energy dissipation, are such that

$$e_1^\top \mathcal{R}_1(x_1, x_2, e_1) \geq 0, \quad e_2^\top \mathcal{R}_2(x_1, x_2, e_1) \geq 0 \tag{4}$$

Cyclo-passivity is confirmed by the computation of the following *differential dissipation inequality*

$$\begin{aligned} \frac{d}{dt}H &= y_1^\top u_1 + y_2^\top u_2 \\ &\quad - e_1^\top \mathcal{R}_1(x_1, x_2, e_1) - e_2^\top \mathcal{R}_2(x_1, x_2, e_1) \\ &\leq y_1^\top u_1 + y_2^\top u_2 \end{aligned} \tag{5}$$

Indeed, time-integration of the differential inequality and substitution of equality of initial and final state implies (1). Furthermore we note that if H is bounded from below then in fact the system is *passive*.

The assumptions reflected in the special structure of the port-Hamiltonian system (2) mean that the interaction between the x_1 part and the x_2 part of the system takes place only through the Hamiltonian $H(x_1, x_2)$, and *not* via a network coupling (the off-diagonal blocks of the J-matrix are zero).

Clearly, the assumed form of the equations implies that for any initial condition and any input function u_2 there exists an input function u_1 that keeps x_1 equal to a constant value $\bar{x}_1$. It follows, cf. Van der Schaft, Jeltsema (2021), that for any such input functions the system is cyclo-passive at port 2 with storage function $H(\bar{x}_1, x_2)$, and cyclo-passive at port 1 as well, with zero storage function.

Furthermore, instead of keeping x_1 constant, u_1 may be also chosen such as to keep $y_1 = e_1$ equal to some constant value $\bar{y}_1$. In this case it follows, cf. Van der Schaft, Jeltsema (2022, 2021), that the system for any such input function is again cyclo-passive at port 1 and port 2. In fact, the storage function for cyclo-passivity at port 1 is given as $\bar{y}_1^\top x_1$, and at port 2 by minus the partial Legendre transform of H with respect to y_1, that is

$$H_1^*(\bar{y}_1, x_2) = H(x_1, x_2) - \bar{y}_1^\top x_1, \quad y_1 = \frac{\partial H}{\partial x_1}(x_1, x_2), \quad (6)$$

where it is assumed that x_1 can be solved from the second equation as a function of y_1, x_2. (The * notation refers to the Legendre transform, while the subscript $_1$ is used because the *partial* Legendre transform with respect to x_1 is taken.) Indeed, in view of the assumptions on the structure of the system, we obtain

$$\frac{d}{dt} H_1^*(\bar{y}_1, x_2) \leq y_2^\top u_2 \quad (7)$$

This latter property was called *one-port cyclo-passivity* in Van der Schaft, Jeltsema (2022). Clearly, both cyclo-passivity properties imply restrictions on the possibilities for converting energy at port 1 to port 2: this is *not* possible in a recurrent manner while keeping either x_1 or y_1 constant. In general, such restrictions are not present in multiphysics systems whose port-Hamiltonian formulation exhibit non-zero off-diagonal terms in the J-matrix; see the analysis in Van der Schaft, Jeltsema (2022, 2021).

Note that the property of one-port cyclo-passivity in the thermodynamic case is exactly the same as the classical observation by Carnot that no thermal energy can be converted into mechanical energy ('work') while keeping the temperature of the heat source constant. Said otherwise, thermodynamic systems are one-port cyclo-passive (with first port being the thermal port and second port the mechanical port), but there are many other systems that share this property; see Van der Schaft, Jeltsema (2022, 2021) for a number of applications in various physical domains.

In this talk we will further elaborate on both notions of cyclo-passivity at one of the ports separately. First, we show that in some cases one-port cyclo-passivity can be *enforced* by adding extra state variables. In particular this shows that one-port cyclo-passivity is strictly speaking *not* an input-output property. This will be illustrated by the example of a linear DC-motor. In this case, the off-diagonal elements of the interconnection structure J between the mechanical and electrical port are non-zero (and in fact given by the gyration constant of the DC-motor). However, by adding the *angle* of the rotor as an extra state variable, a state space transformation can be defined that *eliminates* the coupling given by the gyration constant in the extended (non-minimal) state space formulation. The ramifications of this surprising observation will be investigated, as well connections with related phenomena such as the Blondel-Park transformation of the synchronous generator model. Second, we focus on extensions to the Carnot cycle for one-port cyclo-passive port-Hamiltonian systems. In particular, while the classical Carnot cycle switches between isothermals (y_1 constant), and adiabatics (x_1 constant), one may also consider other cyclic trajectories (involving multiple values of y_1) as well. Furthermore, the Carnot efficiency is defined as the delivered mechanical work divided by the supplied thermal energy during the isothermal at high temperature. This means that the thermal energy (Heat) released during the isothermal at low temperature is sought to be minimized. We will investigate how to define alternative notions of efficiency.

Third, we return to the original statement of the Second Law of thermodynamics. The formulation given by Lord Kelvin states that (see Fermi (1936)):

A transformation of a thermodynamic system whose only final result is to transform into work heat extracted from a source which is at the same temperature throughout is impossible.

Note that this allows for interactions with heat sources at more than one temperature; however such that the net heat that is taken from the sources at different temperatures during the cyclic process is zero; see Fermi (1936); Van der Schaft (2021). In fact, such interactions with multiple heat sources are indispensable in the derivation of Clausius' inequality leading to the definition of entropy. We will take a closer look at the consequences of this for general cyclo-passive port-Hamiltonian systems of the form described in (2).

REFERENCES

E. Fermi, *Thermodynamics*, Prentice-Hall, 1937 (Dover edition 1956).

D. Kondepudi, I. Prigogine. *Modern Thermodynamics; From Heat Engines to Dissipative Structures*, 2nd edition. Wiley, 2015.

A.J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, 3rd Edition 2017, Springer International.

A.J. van der Schaft, Cyclo-dissipativity revisited, *IEEE Transactions on Automatic Control*, 66(6), 2920–2924, 2020.

A.J. van der Schaft, Classical dynamics revisited: a systems and control perspective, *IEEE Control Systems Magazine*, 41(5), 32–60, 2021.

A.J. van der Schaft, D. Jeltsema, "Port-Hamiltonian Systems Theory: An Introductory Overview," *Foundations and Trends in Systems and Control*, 1(2/3), 173–378, 2014.

A.J. van der Schaft, D. Jeltsema, Limits to energy conversion, *IEEE Transactions on Automatic Control*, 67(1), 532–538.

A.J. van der Schaft, D. Jeltsema, On energy conversion in port-Hamiltonian systems, Proc. 60th CDC, Austin, 2021, DOI: 10.1109/CDC45484.2021.9683292

J.C. Willems, Dissipative dynamical systems, Part I: General theory, *Arch. Rat. Mech. and Analysis*, 45(5), 321–351, 1972.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Online parameter tracking in human reaching adaptation and control ⋆

Frédéric Crevecoeur *

* *Institute of Information and Communication Technologies, Electronics and Applied Mathematics, and Institute of Neuroscience, UCLouvain (e-mail: frederic.crevecoeur@uclouvain.be).*

Abstract: Recent experiments have suggested that the nervous system adapted feedback control strategies during an ongoing, perturbed movement. These findings raised the possibility that a function of motor adaptation could be to complement feedback control online, but this idea had not been tested with biologically realistic properties of the human motor system, considering in particular the non-linear limb dynamics and the presence of transmission delays in the neural feedback loop. This study addresses this question by showing that online adaptive control is indeed feasible in a simplified nonlinear model of the human arm, featuring a delay of 60ms as observed in experiments. It is shown that online adaptation can reduce the impact of non-linear effects arising due to limb dynamics within a single movement. Strikingly, the directions that most benefited from online adaptation correlated with known directional biases characterising the distribution of reaching representations in the primate's brain. Further, it is demonstrated that, for some movement directions, it is possible to learn to produce relatively straight hand paths with end-point errors comparable with human performance within tens of trials. These simulation results provide support to the hypothesis that a function of adaptation in the human sensorimotor system is to compensate online for unmodelled disturbances arising in novel or non-linear environments.

Keywords: Model-based control; Adaptive neural control; Human reaching; Control in neuroscience.

1. INTRODUCTION

A paradigmatic example of adaptation in biological control is the ability of primates to learn to compensate for force fields applied during reaching movements. Although it is often assumed that adaptation improves or preserves motor performances across movements (Shadmehr et al. (2010)), it has been recently suggested that adaptation of human reaching control was fast enough to influence an ongoing movement (Mathew and Crevecoeur (2021)).

These previous findings were modelled in the framework of adaptative optimal control (Bitmead et al. (1990)): a real-time identification procedure adapting the model parameters of a Linear Quadratic Gaussian regulator (LQG) online. This approach was applied to a simplified, linear, and fully observable model of limb control. It remained unknown whether this interpretation was amenable to more realistic models of reaching movements, including non-linear mechanical effects, noise, and sensorimotor delays.

The present study addresses this question by showing that it is possible based on linear approximations to control a human-inspired, non-linear two jointed arm with noise and temporal delays consistent with the sensorimotor system of humans (∼60ms, Scott (2016)). Moreover, it is shown that the benefits of the adaptive control mostly impacted reaching trajectories in directions that broadly corresponded to known directional biases in the distribution of movement representations in monkey's primary motor cortex. In all, the simulation results supported the hypothesis that the human nervous system adapts closed-loop control of reaching movements online.

⋆ FC is supported by a grant from F.R.S.-FNRS (Belgium, 1.C.033.18).

2. ADAPTIVE CONTROL MODEL

2.1 Mechanical Model

The biomechanical system is described by the following differential equation: let $\theta = [\theta_s, \theta_e]^T$ denote the vector of shoulder and elbow angles (θ_s and θ_e, respectively), $\tau = [\tau_s, \tau_e]$ the vector of shoulder and elbow torques (Fig. 1), the second order differential equation governing the movement in the horizontal plane is (Li and Todorov (2007)):

$$\ddot{\theta} = M^{-1}(\theta)\left(\tau - C(\theta, \dot{\theta}) - D\dot{\theta}\right), \tag{1}$$

where $M(\theta) \in \mathbf{R}^{2\times 2}$ is the matrix of moments of inertia that depends on the cosine of θ_e, $C(\theta, \dot{\theta}) \in \mathbf{R}^2$ is a vector of nonlinear effects dependent on the factors $\sin(\theta_2)$, $\dot{\theta}_s\dot{\theta}_e$, $\dot{\theta}_s^2$, and $\dot{\theta}_e^2$, and $D \in \mathbf{R}^{2\times 2}$ captures linear viscous forces opposing to velocity (see Appendix).

The mechanical system is coupled with a linear model of muscles dynamics corresponding to a first-order, low-pass filter with time constant $\delta = 60ms$. A stochastic

disturbance is added to this model to include additive noise impacting motor commands (Jones et al. (2002)):

$$d\tau = \frac{1}{\delta}\left(u - \tau\right)dt + \alpha dW, \tag{2}$$

with $u \in \mathbf{R}^2$ representing the command vector, α is a scaling parameter and dW captures instantaneous changes of a standard Brownian motion.

2.2 Controller Definition

The design of an LQG controller for human reaching control follows previous work (Todorov and Jordan (2002)). The linearised model of (1) used in the controller is derived around an equilibrium point corresponding to the starting location of a movement θ_0, and by setting angular velocities to zero:

$$\ddot{\theta} \simeq M^{-1}(\theta_0)\left(\tau - D\dot{\theta}\right). \tag{3}$$

The system is discretised and coupled with (2) with explicit Euler integration (time step of $\Delta t = 10ms$). A static goal target (θ^*, with $\dot{\theta}^* = 0$) is added to the state vector, and the difference equation becomes:

$$x_{k+1} = Ax_k + Bu_k + \xi_k, \tag{4}$$

with $x = [\theta_s, \theta_e, \dot{\theta}_s, \dot{\theta}_e, \tau_s, \tau_e, \theta_s^*, \theta_e^*]^T$, and ξ_k being zero-mean, multivariate Gaussian disturbance with known covariance matrix Ω_ξ, coming from i.i.d. increments of αdW over one time step. The discrete time difference equation is coupled with an output measurement signal that included the delay in the feedback loop:

$$y_k = Hx_{k-h} + \omega_k, \tag{5}$$

with ω_k being zero-mean, Gaussian noise with known covariance matrix Ω_ω, and h representing the delay in number of sample times. This measurement signal represents the sensory data that is available to the brain, it is assumed that joint angles, velocities and actuator forces are encoded in limb afferent feedback. The delay is handled with system augmentation. We defined $z_k^T := [x_k^T, x_{k-1}^T, \ldots, x_{k-h}^T]$ and we used $h = 6$, compatible with the physiological long-latency delay of 60ms (Scott (2016)). The augmented state-space model matrices denoted by $\bar{A}$, $\bar{B}$ and $\bar{H}$ are defined in the Appendix. The state estimator is based on a predictive Kalman filter following standard definition:

$$\hat{z}_{k+1} = \bar{A}\hat{z}_k + \bar{B}u_k + K_k\left(y_k - \bar{H}\hat{z}_k\right). \tag{6}$$

Regarding movement simulations, we used finite horizon formulation with $N = 60$ time steps ($600ms$). The cost-function is a standard quadratic form defined as follows:

$$J(z,u) = z_N^T Q_N z_N + \sum_{k=1}^{N-1} z_k^T Q_k z_k + u_k^T R u_k. \tag{7}$$

We used a kinematic constraint defined on the change in joint angles during the first 10 time steps to penalise deviations from a straight line. Such kinematic constraint has been suggested in previous models of human reaching movements and simply captures the fact that humans spontaneously tend to reach straight (Mistry et al. (2013)). The other matrices for the running costs ($Q_k, 10 < k < N$) were set to zero. The terminal constraint was defined such that:

$$z_N^T Q_N z_N = w_1\|\theta_N - \theta^*\|^2 + w_2\|\dot{\theta}_N\|^2, \tag{8}$$

with θ^* the target joint coordinate and w_1 and w_2 are parameters. Their values were set manually to $w_1 = 500$, and $w_2 = 10$, and $R = 10^{-5}\mathbf{I}_{2\times 2}$. Based on these definitions, a standard LQG including an optimal linear state-feedback controller and optimal Kalman gains could be derived (Astrom (1970)). The next section presents how this controller was updated through time based on system identification.

2.3 Adaptive State-Feedback Control

The previous section set up a linear state-feedback controller for a non-linear system, resulting in model errors induced by non-linear effects. The adaptive optimal control model was based on the idea that the parameters of the linear model could be updated over time to compensate locally for non-linear effects. Schematically, the LQG controller featured a Kalman filter (Fig. 1, K) and a linear state-feedback controller that mapped the estimated state into motor commands (Fig. 1, C). In parallel, the error between the expected and measured feedback could be used to update the system model (Fig. 1, ID). This paper used recursive least square identification (Bitmead et al. (1990)). Calling Θ_{k-1} the estimated set of parameters at time $k-1$, the prediction error obtained at time k is:

$$\epsilon_k(\Theta_{k-1}) = y_k - \bar{H}\hat{z}_k \tag{9}$$

$$= y_k - H\hat{x}_{k-h}. \tag{10}$$

Observe that the second term of the right-hand side of (9) is the expected system output. Since it depends on the parameters through the use of the system matrices A and B in the predictive Kalman filter, we introduce the notation: $\hat{y}_k(\Theta_{k-1}) := \bar{H}\hat{z}_k$. The parameter update is:

$$\Theta_k = \Theta_{k-1} + \gamma\frac{\partial\hat{y}_k(\Theta_{k-1})}{\partial\Theta}\epsilon_k, \tag{11}$$

where γ is the online learning rate. In all the controller could be summarised as follows:

- A linear state-feedback controller dependent on estimated parameters $u_k := L(\Theta_k)\hat{x}_k$,
- An update of the estimated parameters Θ_k, followed by a re-computation of optimal control and Kalman gains with the novel parameter estimate.

In practice the set of parameters that updated was constrained by observing that the non-linearities impacted the relationship between torques and joint accelerations, in other words only the entries of the matrix A that approximated $M^{-1}(\theta)$ and $C(\theta,\dot{\theta})$ were allowed to change over time.

2.4 Heuristic Rules

Our first results below show the effect of the combined control and identification to compensate for non-linear effects online. Then, we explored in simulations whether

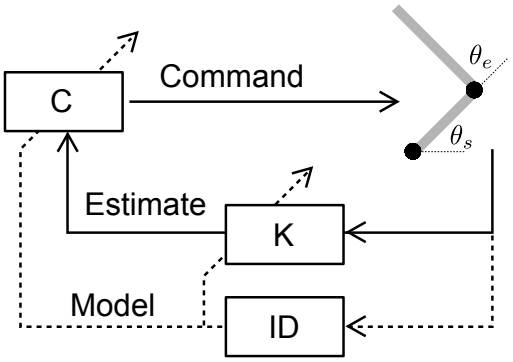


Fig. 1. Schematic depiction of the adaptive optimal feedback controller: (C) controller, (K) Kalman filter, (ID) system identification procedure, which updates the model parameters used in the state estimator and in the controller (dashed arrows). Solid arrows are the state feedback controller corresponding to the long-latency feedback loop.

it was possible to learn a representation of limb dynamics following repeated practice of the same movement. We used the following update rules defined heuristically but compatible with human behaviour:

(1) Gradual adaptation requires that the model updates carried over to the next trial. To limit the impact of noise, we computed for each trial an initial set of parameters corresponding to the average of the previous movement. We denote the time-varying parameters of trial i by $\Theta_{k,i}$ and the time average over a trial by $\bar{\Theta}_{.,i}$.
(2) We defined a composite cost in Cartesian coordinates to capture humans' spontaneous tendency to reach straight and to reduce end-point errors. This cost was the sum of the square end-point error ($d_e^2(i)$) and the square maximum perpendicular displacement relative to a straight line ($d_o^2(i)$), for the i^{th} movement. Let $c(i) := d_e^2(i) + d_o^2(i)$, we set $\Theta_{1,i+1}$ to $\bar{\Theta}_{.,i}$ if the condition $c(i) \leq \varepsilon c(i-1)$ was true with $\varepsilon = 1.05$. In other words, if the identification disrupted the trajectory such that the composite cost increased by more than 5%, due to noise or to an inadequate value of γ, the corresponding changes in estimated parameters were not retained.
(3) The online learning rate was adjusted from trial-to-trial: it was either divided in half if $c(i) > \varepsilon c(i-1)$, or multiplied by two (with the initial rate as upper limit).

3. RESULTS

Two standard tasks were simulated. The first task was a centre-out reaching task towards targets placed along a circle of radius $10cm$ around the start position corresponding to $45deg$ and $90deg$ of shoulder and elbow angles (Fig. 2a). Because the linearisation was computed around the starting point with zero velocity, it is expected that the trajectories are perturbed by non-linear effects away from this equilibrium, that is for movements of large amplitude or with high velocity. In particular, movements including the shoulder angle produce non-linear effects due to the function $C(\theta, \dot{\theta})$. It can be observed that these movements, mostly located in the second and fourth quadrant of the

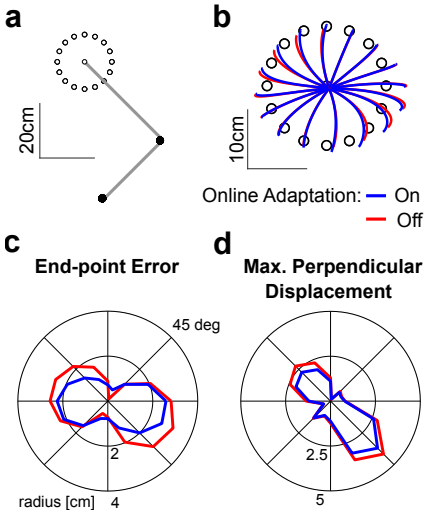


Fig. 2. Centre-out reaching task: (a) joint configuration with the start target at $\theta_s = 45$deg and $\theta_e = 90$deg plus all peripheral targets distants by 10cm; (b) end-point trajectories with (blue) or without (red) online learning; (c) polar representation of the norm of the end-point error aligned with the corresponding target angle; (d) same as (c) with the maximum perpendicular displacement.

workspace (Fig. 2b), were those that mostly benefited from the adaptive estimation of Θ_k. Indeed, with online identification (Fig. 2, blue), it was possible within a single trial to reduce the end-point error as well as the maximum perpendicular displacement (Fig. 2c and d). The simulations of Fig. 2 were obtained without process noise ($\Omega_\xi = \Omega_\omega = 0$), and an online learning rate of $\gamma = 0.005$.

The simulations in Fig. 2 show that it was possible to improve a linear control applied to a nonlinear system within a single reaching movement simply based o the observation that the movements with online adaptation tended to be straighter. Next, it was hypothesised that the controller could further improve if changes in Θ were kept in memory for the next movement using the heuristic learning rules defined above. It was found for the selected movement direction and amplitude (which includes a large change in shoulder angle, Fig. 3a) that iteratively updating the linear model across trials yielded movement paths that were relatively straight and accurate, with end-point error and maximum perpendicular displacements in the range of those observed in experiments ($\leq$ 2cm, Fig. 3b and c). These simulations included noise and an initial learning rate set manually to $\gamma = 7 \times 10^{-5}$. The performance of the control and estimation algorithm plateaued after ~ 30 trials, when the online learning rate γ was very close to 0, suggesting it was close to a local minimum for this particular movement, linearization, and parameter set.

4. DISCUSSION

The simulations showed that online identification of the parameters of a linear model of reaching control could at least partially compensate for non-linear effects in the presence of noise and delays compatible with the sensorimotor system of humans. The improvements could be observed within a single movement, as well as across few

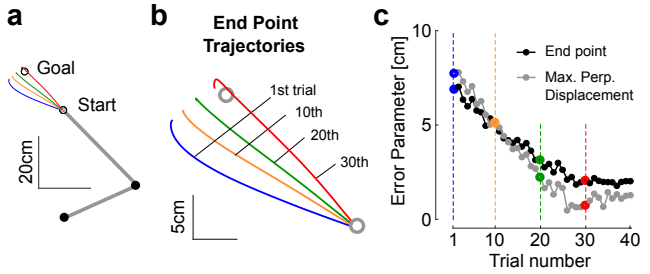


Fig. 3. (a) Joint configuration, start, and goal targets; (b) Selected movement traces at different stages of the learning process; (c) End-point error and maximum perpendicular displacement.

tens of trials when consecutive estimates were averaged and reused iteratively. The foregoing results thus provided support to the hypothesis that a function of sensorimotor adaptation is not only to adapt movement control over medium or long timescales, but also to complement feedback control online.

Strikingly, because the largest impact of the model non-linearities arose in directions that included larger amount of shoulder motion, the benefits of the adaptive controller were mostly visible in the second and fourth quadrant of the workspace, which reproduced known biases in reach representation observed in the primate's brain (Lillicrap and Scott (2013)). This previous paper showed that the directional biases emerged from limb biomechanics, since they were reproduced in an artificial system featuring the details of limb and muscles dynamical properties. The present results propose an complementary view on the reason why whole limb flexion and extension movements may be over-represented in primary motor cortex: it is conceivable that an adaptive neural feedback controller needs more resources in the directions where non-linearities of the limb have a larger impact.

As concluding remarks, two important aspects must be emphasised. First, the combination of adaptation and control, and the iterative improvement across consecutive movements, were obtained without using any feedforward pathway. Thus, the results warrant further investigation to study if or when an often assumed feedforward controller in the brain is actually required. Second, the model can only improve: the online learning rate was not optimised, the identification procedure was not tuned to filter out the noise, the linear representation was low-dimensional (and derived from first principles), and the heuristic adaptation rules were simplistic. Optimising each component is expected to make the adaptive state-feedback controller a powerful candidate model of human reaching control.

ACKNOWLEDGEMENTS

The author want to thank J. Hendrickx for critical comments on an earlier version of the abstract.

REFERENCES

Astrom, K. (1970). *Introduction to Stochastic Control Theory*. Academic Press, New York, London.

Bitmead, R., Gevers, M., and Wertz, V. (1990). *Adaptive optimal control : the thinking man's GPC*. Prentice Hall, New York London Toronto [etc].

Crevecoeur, F. and Scott, S. (2014). Beyond muscle stiffness: importance of state estimation to account for very fast motor corrections. *PLoS Computational Biology*, 10, e1003869.

Jones, K., Hamilton, A., and Wolpert, D. (2002). Sources of signal-dependent noise during isometric force production. *Journal of Neurophysiology*, 88, 1533–1544.

Li, W. and Todorov, E. (2007). Iterative linearization methods for approximately optimal control and estimation of non-linear stochastic system. *International Journal of Control*, 80, 1439–1453.

Lillicrap, T. and Scott, S. (2013). Preference distributions of primary motor cortex neurons reflect control solutions optmimzed for limb biomechanics. *Neuron*, 77, 168–179.

Mathew, J. and Crevecoeur, F. (2021). Adaptive feedback control in human reaching adaptation to force fields. *Frontiers in Human Neuroscience*, 15, 742608.

Mistry, M., Theodorou, E., Schaal, S., and Kawato, M. (2013). Optimal control of reaching includes kinematic constraints. *Journal of Neurophysiology*, 110, 1–11.

Scott, S. (2016). A functional taxonomy of bottom-up sensory feedback processing for motor actions. *Trends in Neurosciences*, 39, 512–526.

Shadmehr, R., Smith, M., and Krakauer, J. (2010). Error correction, sensory prediction and adaptation in motor control. *Annual Review of Neuroscience*, 33, 89–108.

Todorov, E. and Jordan, M. (2002). Optimal feedback control as a theory of motor coodination. *Nature Neuroscience*, 5, 1226–1235.

Appendix A. DEFINITION OF MODEL PARAMETERS

The definitions and numerical values followed Li and Todorov (2007) and Crevecoeur and Scott (2014):

$$M(\theta) = \begin{pmatrix} a_1 + 2a_2\cos(\theta_e) & a_3 + a_2\cos(\theta_e) \\ a_3 + a_2\cos(\theta_e) & a_3 \end{pmatrix}, \quad \text{(A.1)}$$

$$C(\theta, \dot{\theta}) = \begin{pmatrix} -\dot{\theta}_e(2\dot{\theta}_s + \dot{\theta}_e) \\ \dot{\theta}_s^2 \end{pmatrix} a_2 \sin(\theta_e), \quad \text{(A.2)}$$

$$D = \begin{pmatrix} 0.14 & 0.014 \\ 0.014 & 0.14 \end{pmatrix}, \quad \text{(A.3)}$$

and the a_i were parameters linked to the segment masses and inertias. The numerical values were: $a_1 = 0.28$, $a_2 = 0.048$, and $a_3 = 0.1kgm^2$. The matrix H was the identity. The augmented matrices to include de delay were defined as follows:

$$\bar{A} = \begin{pmatrix} A & 0 & \dots & 0 \\ I & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & I & 0 \end{pmatrix}, \bar{B} = \begin{pmatrix} B \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{(A.4)}$$

$$\bar{H} = \begin{pmatrix} 0 & \dots & 0 & H \end{pmatrix}. \quad \text{(A.5)}$$

The noise covariance matrices were $\Omega_\xi = 0.01\bar{B}\bar{B}^T$ such that the noise only affected the control signal, and $\Omega_\omega = 0.01I_{8\times 8}$. The latter definition considered the possibility that all sensory signals were corrupted by additive noise.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Data-driven dissipativity analysis of linear systems

Henk J. van Waarde * **M. Kanat Camlibel** *
Paolo Rapisarda ** **Harry L. Trentelman** *

* *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands (h.j.van.waarde@rug.nl, m.k.camlibel@rug.nl, h.l.trentelman@rug.nl).*
** *Vision, Learning and Control group, University of Southampton, United Kingdom, (pr3@ecs.soton.ac.uk)*

Abstract: The concept of dissipativity is a cornerstone of systems and control theory. Typically, dissipativity properties are verified by resorting to a mathematical model of the system under consideration. In this extended abstract, we aim at assessing dissipativity by computing storage functions for linear systems directly from measured data. As our main contributions, we provide conditions under which dissipativity can be ascertained from a finite collection of noisy data samples. Different noise models will be considered that can capture a variety of situations, including the cases that the data samples are noise-free, the energy of the noise is bounded, or the individual noise samples are bounded. All of our conditions are phrased in terms of data-based linear matrix inequalities, which can be readily solved using existing software packages.

Keywords: System identification, dissipativity, linear systems.

1. INTRODUCTION

As is generally acknowledged, the concept of dissipativity (Willems, 1972a), (Willems, 1972b) has formed the foundation for large parts of systems and control theory as developed in the past fifty years. Indeed, the above papers together with Willems' work on linear quadratic problems and the associated algebraic Riccati equation (Willems, 1971) are generally considered to provide the main concepts and analysis tools in many areas of linear and nonlinear systems and control, ranging from stability theory, linear quadratic optimal control and stochastic realization theory, to network synthesis, differential games and robust control.

In the present work, we study dissipativity of linear finite-dimensional input-state-output systems from a data-driven perspective. It is well-known that for a given input-state-output system with given supply rate, one can test dissipativity by checking the feasibility of a linear matrix inequality involving the system matrices. In this extended abstract, we assume that the system dynamics are not known a priori. In this situation, the question arises whether we can verify dissipativity using measured system trajectories, instead of a system model.

Recently, the problem of inferring dissipativity properties from data has received considerable attention. The most relevant references for this extended abstract are (Maupong et al., 2017; Romer et al., 2019; Koch et al., 2020a; Steentjes et al., 2021). In (Maupong et al., 2017), the notion of (finite-horizon) L-dissipativity was introduced and also studied in (Romer et al., 2019). A discrete-time system is L-dissipative if the average of the supply rate over the interval $[0, L]$ is nonnegative for all system trajectories. In both these contributions, a crucial assumption is that the input trajectory is persistently exciting of a sufficiently high order (see (Willems et al., 2005) and (van Waarde et al., 2020a)). This property of the input sequence can be shown to imply that the data-generating system is uniquely identifiable from the data.

In this extended abstract we adopt the more classical notion of dissipativity for linear systems, rather than L-dissipativity, similar to the setup of (Koch et al., 2020a), see also (Steentjes et al., 2021) for dissipativity analysis and controller synthesis of interconnected networks. Our aim is to provide necessary and sufficient conditions for dissipativity based on data, for noiseless and noisy measurements.

Our approach involves bounding the noise by a quadratic matrix inequality, which implies that also the unknown system parameters satisfy a quadratic matrix inequality. Our goal is to ascertain dissipativity of *all* systems satisfying this inequality. The method thus fits in the robust control literature, where quadratic uncertainty descriptions have been studied in detail. We mention contributions to integral quadratic constraints (Megretski and Rantzer, 1997), the quadratic separator (Iwasaki and Hara, 1998), and the full block S-procedure (Scherer, 1997, 2001). We will apply a matrix S-lemma (van Waarde et al., 2022) that is a recent extension to matrix variables of the famous S-lemma (Yakubovich, 1977).

The outline of the extended abstract is as follows. In Section 2 we give a short recap of the concept of dissipativity of linear input-state-output systems. Then, in Section 3 we state the problem of data-driven dissipativity. Section 4 contains our results. First, we show that dissipativity of

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

an unknown linear system can only be ascertained on the basis of the given data if a matrix constructed from measured states and inputs has full rank. In the noiseless data case, this implies that one can only verify dissipativity from data if the data-generating system is the only one that explains the data, in other words, if the true system is identifiable from the data. In this case, dissipativity of the unknown system can be ascertained by checking the feasibility of a given data-based linear matrix inequality. In the noisy data case, it turns out that one does not need identifiability. In order to check dissipativity in this case, we combine the matrix S-lemma with a basic dualization lemma to provide a data-driven test for dissipativity. Finally, Section 5 contains our conclusions.

Notation

The *inertia* of a symmetric matrix S is denoted by $\mathrm{In}(S) = (\rho_-, \rho_0, \rho_+)$ where ρ_-, ρ_0, and ρ_+ respectively denote the number of negative, zero, and positive eigenvalues of S.

2. DISSIPATIVITY OF LINEAR SYSTEMS

Consider a linear discrete-time input-state-output system

$$\boldsymbol{x}(t+1) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \tag{1a}$$
$$\boldsymbol{y}(t) = C\boldsymbol{x}(t) + D\boldsymbol{u}(t) \tag{1b}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$.

Let $S = S^\top \in \mathbb{R}^{(m+p)\times(m+p)}$. The system (1) is said to be *dissipative* with respect to the *supply rate*

$$s(u,y) = \begin{bmatrix} u \\ y \end{bmatrix}^\top S \begin{bmatrix} u \\ y \end{bmatrix} \tag{2}$$

if there exists $P \in \mathbb{R}^{n\times n}$ with $P = P^\top \geqslant 0$ such that the *dissipation inequality*

$$\boldsymbol{x}(t)^\top P \boldsymbol{x}(t) + s\big(\boldsymbol{u}(t), \boldsymbol{y}(t)\big) \geqslant \boldsymbol{x}(t+1)^\top P \boldsymbol{x}(t+1) \tag{3}$$

holds for all $t \geqslant 0$ and for all trajectories $(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) : \mathbb{N} \to \mathbb{R}^{m+n+p}$ of (1).

It follows from (3) that dissipativity with respect to the supply rate (2) is equivalent to the feasibility of the linear matrix inequalities $P = P^\top \geqslant 0$ and

$$\begin{bmatrix} I & 0 \\ A & B \end{bmatrix}^\top \begin{bmatrix} P & 0 \\ 0 & -P \end{bmatrix} \begin{bmatrix} I & 0 \\ A & B \end{bmatrix} + \begin{bmatrix} 0 & I \\ C & D \end{bmatrix}^\top S \begin{bmatrix} 0 & I \\ C & D \end{bmatrix} \geqslant 0. \tag{4}$$

3. PROBLEM FORMULATION

Consider the linear discrete-time input-state-output system

$$\boldsymbol{x}(t+1) = A_s\boldsymbol{x}(t) + B_s\boldsymbol{u}(t) + \boldsymbol{w}(t) \tag{5a}$$
$$\boldsymbol{y}(t) = C_s\boldsymbol{x}(t) + D_s\boldsymbol{u}(t) + \boldsymbol{z}(t) \tag{5b}$$

where $(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{m+n+p}$ are the input, state and output, and $(\boldsymbol{w}, \boldsymbol{z}) \in \mathbb{R}^{n+p}$ are noise terms. Throughout the extended abstract, we assume that the "true" system matrices (A_s, B_s, C_s, D_s) and the noise $(\boldsymbol{w}, \boldsymbol{z})$ are *unknown*. What is known instead are a finite number of input-state-output measurements of (5), which we collect in the matrices

$$U_- := [u(0)\ u(1)\ \cdots\ u(T-1)]$$
$$X := [x(0)\ x(1)\ \cdots\ x(T)]$$
$$Y_- := [y(0)\ y(1)\ \cdots\ y(T-1)].$$

We will also make use of the auxiliary matrices

$$X_- := [x(0)\ x(1)\ \cdots\ x(T-1)]$$
$$X_+ := [x(1)\ x(2)\ \cdots\ x(T)].$$

The goal of this extended abstract is to infer dissipativity properties of the true system from the data (U_-, X, Y_-).

We define

$$\Sigma^{\mathcal{N}} = \left\{ (A,B,C,D) \mid \begin{bmatrix} X_+ \\ Y_- \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X_- \\ U_- \end{bmatrix} \in \mathcal{N} \right\},$$

where $\mathcal{N} \subseteq \mathbb{R}^{(n+p)\times T}$ is a set defining a *noise model* to be specified below. We assume that

$$(A_s, B_s, C_s, D_s) \in \Sigma^{\mathcal{N}}. \tag{7}$$

In the sequel, we will consider two types of noise models. The first one will capture noise-free situations in which the measurements (U_-, X, Y_-) are exact:

$$\mathcal{N}_0 := \{0\}. \tag{8}$$

The second noise model is defined by

$$\mathcal{N}_1 := \left\{ V \in \mathbb{R}^{(n+p)\times T} \mid \begin{bmatrix} I \\ V^\top \end{bmatrix}^\top \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{12}^\top & \Phi_{22} \end{bmatrix} \begin{bmatrix} I \\ V^\top \end{bmatrix} \geqslant 0 \right\}, \tag{9}$$

where $\Phi_{11} = \Phi_{11}^\top \in \mathbb{R}^{(n+p)\times(n+p)}$, $\Phi_{12} \in \mathbb{R}^{(n+p)\times T}$, and $\Phi_{22} = \Phi_{22}^\top \in \mathbb{R}^{T\times T}$ are known matrices. This noise model was studied before (van Waarde et al., 2022) in the context of data-driven quadratic stabilization and H_2 and H_∞ control. In order to be able to discuss some special cases of the noise model (9), we label the columns of V as $[v(0)\ v(1)\ \cdots\ v(T-1)]$. In the special case $\Phi_{12} = 0$ and $\Phi_{22} = -I$, the bound in (9) reduces to

$$VV^\top = \sum_{t=0}^{T-1} v(t)v(t)^\top \leqslant \Phi_{11}. \tag{10}$$

This inequality can be interpreted as an energy bound on the noise. In addition, norm bounds on the individual noise samples $v(t)$ also give rise to bounds of the form (10), although this generally leads to some conservatism. Indeed, note that $\|v(t)\|^2 \leqslant \epsilon$ implies that $v(t)v(t)^\top \leqslant \epsilon I$ for all t. As such, the bound (10) is satisfied for $\Phi_{11} = T\epsilon I$. We also note that $\mathcal{N}_1$ can be interpreted as the dual model to the one studied in (Berberich et al., 2020), and both models are equivalent under mild conditions (van Waarde et al., 2021).

We now define the property of *informativity for dissipativity*, which is the main concept studied in this extended abstract. This definition is inspired by (van Waarde et al., 2020b), and we refer to that paper for a general treatment.

Definition 1. Let a noise model $\mathcal{N}$ be given. The data (U_-, X, Y_-) are *informative for dissipativity* with respect to the supply rate (2) if there exists a $P = P^\top \geqslant 0$ such that (4) holds for every system $(A, B, C, D) \in \Sigma^{\mathcal{N}}$.

The rationale behind Definition 1 is as follows: on the basis of the given data we are unable to distinguish between the systems in $\Sigma^{\mathcal{N}}$ in the sense that any of these systems could have generated the data. Nonetheless, if *all* of these systems are dissipative, then we can also conclude that the *true* data-generating system is dissipative. Note that we restrict our attention to the situation in which the systems in $\Sigma^{\mathcal{N}}$ are dissipative with a *common* storage function.

The following assumptions will be valid throughout:

(A1) The matrix S has inertia $\mathrm{In}(S) = (p, 0, m)$.
(A2) The set $\mathcal{N}_1$ is bounded and has nonempty interior.

It is a well-known fact that a necessary condition for dissipativity of any system of the form (1) is that $m \leqslant \rho_+$, i.e., the input dimension does not exceed the positive signature of S. Assumption (A1) requires that the input dimension is equal to this positive signature, and in addition that the matrix S is nonsingular. This assumption is satisfied, for example, for the positive-real and bounded-real case (Scherer and Weiland, 1999). Indeed, in the positive-real case we have that $m = p$ and

$$S = \begin{bmatrix} 0 & I_m \\ I_m & 0 \end{bmatrix},$$

so that $\mathrm{In}(S) = (m, 0, m)$. In the bounded-real case we have

$$S = \begin{bmatrix} \gamma^2 I_m & 0 \\ 0 & -I_p \end{bmatrix}$$

for $\gamma > 0$, which implies that $\mathrm{In}(S) = (p, 0, m)$. Assumption (A2) can be verified straightforwardly by assessing certain definiteness properties of the matrix Φ (van Waarde et al., 2021).

The main contribution of this extended abstract is to provide necessary and sufficient conditions for data informativity for the noise models $\mathcal{N}_0$ and $\mathcal{N}_1$.

4. MAIN RESULTS

In this section we state our main results. We will not provide proofs here, but instead refer the interested reader to the extended manuscript (van Waarde et al., 2021).

4.1 A necessary condition for informativity

We begin with a necessary condition for informativity, that applies to both noise models $\mathcal{N}_0$ and $\mathcal{N}_1$.

Theorem 1. Let a noise model $\mathcal{N}$ be given. If the data (U_-, X, Y_-) are informative for dissipativity with respect to the supply rate (2), then

$$\mathrm{rank} \begin{bmatrix} X_- \\ U_- \end{bmatrix} = n + m. \tag{11}$$

Essentially, Theorem 1 and the rank condition (11) formalize the intuition that dissipativity can only be assessed from data that are sufficiently rich. In the noise-free setting (involving model $\mathcal{N}_0$), the rank condition (11) implies that the system matrices A_s, B_s, C_s and D_s can be uniquely identified from the (U_-, X, Y_-)-data. In this setting, the interpretation of Theorem 1 is that dissipativity can *only* be verified from data that are rich enough to uniquely identify the underlying data-generating system.

4.2 Informativity and noiseless data

We now give a characterization of informativity for dissipativity for the noiseless case.

Theorem 2. Consider the noise model $\mathcal{N}_0$. The data (U_-, X, Y_-) are informative for dissipativity with respect to the supply rate (2) if and only if

$$\mathrm{rank} \begin{bmatrix} X_- \\ U_- \end{bmatrix} = n + m \tag{12}$$

and there exists $P = P^\top \geqslant 0$ such that

$$\begin{bmatrix} X_- \\ X_+ \end{bmatrix}^\top \begin{bmatrix} P & 0 \\ 0 & -P \end{bmatrix} \begin{bmatrix} X_- \\ X_+ \end{bmatrix} + \begin{bmatrix} U_- \\ Y_- \end{bmatrix}^\top S \begin{bmatrix} U_- \\ Y_- \end{bmatrix} \geqslant 0. \tag{13}$$

Theorem 2 provides a data-based condition for dissipativity in terms of a linear matrix inequality. Linear matrix inequalities can be solved using standard software packages. We note, however, that such solvers are known to be unreliable for LMI's which define feasible sets without interior points. As such, from a numerical point of view it is desirable that there exists a positive definite P such that left-hand side of (13) is positive definite.

We note that the condition of Theorem 2 has appeared in a similar setting in (Koch et al., 2020b, Thm. 4) and (Koch et al., 2020a, Thm. 3), where an "if"-statement was proven. Our contribution is to prove that these conditions are necessary and sufficient by leveraging Theorem 1.

4.3 Informativity and noisy data

We now consider the noise model $\mathcal{N}_1$ defined in (9). Define

$$N_1 := \left[\begin{array}{c|c} I & \begin{matrix} X_+ \\ Y_- \end{matrix} \\ \hline 0 & \begin{matrix} -X_- \\ -U_- \end{matrix} \end{array}\right] \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{12}^\top & \Phi_{22} \end{bmatrix} \left[\begin{array}{c|c} I & \begin{matrix} X_+ \\ Y_- \end{matrix} \\ \hline 0 & \begin{matrix} -X_- \\ -U_- \end{matrix} \end{array}\right]^\top. \tag{14}$$

We now arrive at the following characterization of informativity for dissipativity, given the noise model $\mathcal{N}_1$.

Theorem 3. Suppose that there exists $V \in \mathbb{R}^{(n+m)\times(n+p)}$ such that

$$\begin{bmatrix} I \\ V \end{bmatrix}^\top N_1 \begin{bmatrix} I \\ V \end{bmatrix} > 0. \tag{15}$$

Partition

$$\begin{bmatrix} \hat{F} & \hat{G} \\ \hat{G}^\top & \hat{H} \end{bmatrix} := -S^{-1},$$

where $\hat{F} = \hat{F}^\top \in \mathbb{R}^{m\times m}$, $\hat{G} \in \mathbb{R}^{m\times p}$, and $\hat{H} = \hat{H}^\top \in \mathbb{R}^{p\times p}$. Given the noise model $\mathcal{N}_1$, the data (U_-, X, Y_-) are informative for dissipativity with respect to the supply rate (2) if and only if there exist a real $n\times n$ matrix $Q = Q^\top > 0$ and a scalar $\alpha \geqslant 0$ such that

$$\begin{bmatrix} Q & 0 & 0 & 0 \\ 0 & \hat{H} & 0 & -\hat{G}^\top \\ 0 & 0 & -Q & 0 \\ 0 & -\hat{G} & 0 & \hat{F} \end{bmatrix} - \alpha N_1 \geqslant 0. \tag{LMI}$$

Theorem 3 provides a tractable method for verifying informativity for dissipativity, given the noise model $\mathcal{N}_1$. The procedure involves solving the linear matrix inequality (LMI) for Q and α. Given Q, a common storage function P for all systems in $\Sigma^{\mathcal{N}_1}$ is readily computable as $P = Q^{-1}$.

The condition (15) implies that the interior of the set of explaining systems $\Sigma^{\mathcal{N}_1}$ is nonempty. The proof of Theorem 3 uses two building blocks. The first one is the so-called matrix S-lemma (van Waarde et al., 2022). This is a generalization to matrix variables of the classical S-lemma, developed in the seventies of the previous century by (Yakubovich, 1977). The second building block is a dualization result that essentially states that the quadruple (A, B, C, D) is dissipative with respect to the supply rate S, with storage function P if and only if the dual

system $(A^\top, C^\top, B^\top, D^\top)$ is dissipative with respect to a related supply rate $\hat{S}$, with storage function P^{-1}, see (van Waarde et al., 2021). A behavioral analogue of this result was obtained in (Willems and Trentelman, 2002, Prop. 12).

We note that a sufficient condition for data-driven dissipativity with a common storage function was given in (Koch et al., 2020a, Thm. 4). The attractive feature of Theorem 3 is that it provides a necessary and sufficient condition, by making use of the matrix S-lemma.

Assuming that the assumptions of Theorem 3 are satisfied, an interesting byproduct on the result is the following: if all systems in $\Sigma^{\mathcal{N}_1}$ are dissipative with common storage function $P = P^\top \geqslant 0$, then P is necessarily *positive definite*. We note that conditions under which all storage functions are positive definite have been studied before in (Hill and Moylan, 1976, Lem. 1), even for nonlinear systems. In that paper, certain minimality conditions were imposed as well as a signature condition on the supply rate. Here, we do not assume minimality but we conclude that all storage functions are positive definite by using Assumption (A1) and an argument related to the noise model.

5. CONCLUSIONS

In this extended abstract we have provided methods to verify dissipativity properties of linear systems directly from measured data. We have considered both the case of exact data and the case that the data are corrupted by noise. In the case of exact data, we have shown that one can only ascertain dissipativity of a system from given data if the system can be uniquely identified from the data. If this is the case, dissipativity can be verified by means of a data-based linear matrix inequality. In the case of noisy data, we have combined the matrix S-lemma (van Waarde et al., 2022) with a dualization property relating dissipativity properties of the original system with those of its dual to characterize data informativity for dissipativity. As in the noiseless case, also in this setting, dissipativity properties of the data-generating system can be ascertained if a data-based LMI is solvable.

REFERENCES

Berberich, J., Koch, A., Scherer, C.W., and Allgöwer, F. (2020). Robust data-driven state-feedback design. In *American Control Conference*, 1532–1538.

Hill, D. and Moylan, P. (1976). The stability of nonlinear dissipative systems. *IEEE Transactions on Automatic Control*, 21(5), 708–711.

Iwasaki, T. and Hara, S. (1998). Well-posedness of feedback systems: insights into exact robustness analysis and approximate computations. *IEEE Transactions on Automatic Control*, 43(5), 619–630.

Koch, A., Berberich, J., and Allgöwer, F. (2020a). Provably robust verification of dissipativity properties from data. *https://arxiv.org/abs/2006.05974*.

Koch, A., Berberich, J., and Allgöwer, F. (2020b). Verifying dissipativity properties from noise-corrupted input-state data. In *Proceedings of the IEEE Conference on Decision and Control*, 616–621.

Maupong, T.M., Mayo-Maldonado, J.C., and Rapisarda, P. (2017). On Lyapunov functions and data-driven dissipativity. *IFAC-PapersOnLine*, 50(1), 7783–7788.

Megretski, A. and Rantzer, A. (1997). System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6), 819–830.

Romer, A., Berberich, J., Köhler, J., and Allgöwer, F. (2019). One-shot verification of dissipativity properties from input-output data. *IEEE Control Systems Letters*, 3(3), 709–714.

Scherer, C.W. (1997). A full block S-procedure with applications. In *Proceedings of the IEEE Conference on Decision and Control*, 2602–2607.

Scherer, C.W. (2001). LPV control and full block multipliers. *Automatica*, 37(3), 361–375.

Scherer, C.W. and Weiland, S. (1999). *Lecture Notes DISC Course on Linear Matrix Inequalities in Control*.

Steentjes, T.R.V., Lazar, M., and Van den Hof, P.M.J. (2021). H_∞-performance analysis and distributed controller synthesis for interconnected linear systems from noisy input-state data. In *Proceedings of the IEEE Conference on Decision and Control*, 3723–3728.

van Waarde, H.J., Camlibel, M.K., Rapisarda, P., and Trentelman, H.L. (2021). Data-driven dissipativity analysis: application of the matrix S-lemma. *https://arxiv.org/abs/2109.02090*.

van Waarde, H.J., Camlibel, M.M., and Mesbahi, M. (2022). From noisy data to feedback controllers: Nonconservative design via a matrix S-lemma. *IEEE Transactions on Automatic Control*, 67(1), 162–175.

van Waarde, H.J., De Persis, C., Camlibel, M.K., and Tesi, P. (2020a). Willems' fundamental lemma for state-space systems and its extension to multiple datasets. *IEEE Control Systems Letters*, 4(3), 602–607.

van Waarde, H.J., Eising, J., Trentelman, H.L., and Camlibel, M.K. (2020b). Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11), 4753–4768.

Willems, J.C. (1971). Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Transactions on Automatic Control*, 16, 621–634.

Willems, J.C. (1972a). Dissipative dynamical systems part I: general theory. *Archive for Rational Mechanics and Analysis*, 45, 321–351.

Willems, J.C. (1972b). Dissipative dynamical systems part II: linear systems with quadratic supply rates. *Archive for Rational Mechanics and Analysis*, 45, 352–393.

Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B.L.M. (2005). A note on persistency of excitation. *Systems & Control Letters*, 54(4), 325–329.

Willems, J.C. and Trentelman, H.L. (2002). Synthesis of dissipative systems using quadratic differential forms - part I. *IEEE Transactions on Automatic Control*, 47, 53–69.

Yakubovich, V.A. (1977). S-procedure in nonlinear control theory. *Vestnik Leningrad University Mathematics*, 4, 73–93.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A constructive proof of the fundamental lemma for data-driven representation of LTI systems ⋆

Ivan Markovsky *

* *Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium; ICREA, Pg. Lluis Companys 23; and CIMNE, Barcelona, Spain*

Eduardo Prieto-Araujo **

** *CITCEA, Universitat Politècnica de Catalunya. Spain. (e-mail: eduardo.prieto-araujo@citcea.upc.edu)*

Florian Dörfler ***

*** *Automatic Control Laboratory (IfA), ETH-Zürich 8092 Zürich, Switzerland (e-mail: dorfler@ethz.ch)*

Abstract: The existing proofs of the fundamental lemma use arguments by contradiction and do not give insight into the assumptions of controllability and persistency of excitation of the input. We present an alternative constructive proof that reduces the required persistency of excitation and characterizes the nongeneric cases in which the extra persistency of excitation beyond the time horizon is needed.

Keywords: behavioral approach; exact identification; persistency of excitation; data-driven control.

1. INTRODUCTION

The fundamental lemma of Willems et al. (2005) addresses the following question: Given a finite trajectory $w_\mathrm{d} \in (\mathbb{R}^q)^T$ of a linear time-invariant (LTI) system $\mathscr{B}$ and a natural number L, $1 \leq L \leq T$, under what conditions the windows of length L

$$\begin{bmatrix} w_\mathrm{d}(1) \\ \vdots \\ w_\mathrm{d}(L) \end{bmatrix}, \begin{bmatrix} w_\mathrm{d}(2) \\ \vdots \\ w_\mathrm{d}(L+1) \end{bmatrix}, \ldots, \begin{bmatrix} w_\mathrm{d}(T-L+1) \\ \vdots \\ w_\mathrm{d}(T) \end{bmatrix},$$

constructed from the length-T trajectory w_d span the space $\mathscr{B}|_L$ of all length-L trajectories of the system?

Using the Hankel matrix

$$\mathscr{H}_L(w_\mathrm{d}) := \begin{bmatrix} w_\mathrm{d}(1) & w_\mathrm{d}(2) & \cdots & w_\mathrm{d}(T-L+1) \\ \vdots & \vdots & & \vdots \\ w_\mathrm{d}(L) & w_\mathrm{d}(L+1) & \cdots & w_\mathrm{d}(T) \end{bmatrix},$$

the question is rephrased as: Given $w_\mathrm{d} \in \mathscr{B}|_T$, where $\mathscr{B}$ is LTI, under what conditions

$$\mathscr{B}|_L = \text{image } \mathscr{H}_L(w_\mathrm{d})? \tag{1}$$

We call (1) a *data-driven representation* of the restricted behavior $\mathscr{B}|_L$. It is used in data-driven analysis, signal processing, and control, see the overview (Markovsky and Dörfler, 2021).

Let m, ℓ, and n be the number of inputs, lag (observability index), and order of $\mathscr{B}$. A necessary and sufficient condition for (1) with $L \geq \ell$ is

$$\text{rank } \mathscr{H}_L(w_\mathrm{d}) = mL + n. \tag{2}$$

This condition is refered to as *generalized persistency of excitation* (Markovsky and Dörfler, 2020). It is verifiable from the data w_d and the prior knowledge of m, ℓ, and n.

Contrary to (2), the solution given in (Willems et al., 2005) assumes a given input/output partitioning $w = \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$ of the variables. Moreover, the fundamental lemma provides sufficient conditions only: (1) holds under the following conditions

A1: $\mathscr{B}$ is controllable and
A2: u_d is persistently exciting of order $L+n$.

The order of persistency of excitation of u_d, denoted $\mathrm{PE}(u_\mathrm{d})$, is the maximal L for which $\mathscr{H}_L(u_\mathrm{d})$ is full row rank. The controllability assumption is not verifiable from the data and is restrictive. In particular, it excludes autonomous systems.

The crux of the fundamental lemma is the need of an extra persistency of excitation of order n in Assumption A2. (Willems et al., 2005) as well as subsequent publications (van Waarde et al., 2020; De Persis and Tesi, 2020; van Waarde et al., 2020; van Waarde, 2021) do not explain why the extra persistency of excitation is needed nor when it is needed. Presently it is not known how conservative assumptions A1 and A2 are.

Initial conditions can be specified by a sufficiently long "past" trajectory $w_\mathrm{d,ini} \in \mathscr{B}|_{T_\mathrm{ini}}$. As shown in (Markovsky and Rapisarda, 2008, Lemma 1), "sufficiently long" is $T_\mathrm{ini} \geq \ell$. Then, the concatenation $w_\mathrm{d,ini} \wedge w_\mathrm{d}$ of $w_\mathrm{d,ini}$ and w_d sets the initial conditions for w_d. Our goal (refered to as "(GOAL)") is to find for given $L \geq \ell$ conditions on u_d and $\mathscr{B}$, under which for any initial condition $w_\mathrm{d,ini}$, a trajectory $w_\mathrm{d,ini} \wedge w_\mathrm{d} \in \mathscr{B}|_{T_\mathrm{ini}+T}$ satisfies (2). Section 2 certify (GOAL) by showing that there is no initial condition for which (2) fails. Assumptions A1 and

⋆ A full version of the paper is submitted to Automatica. Ivan Markovsky received funding from the Catalan Institution for Research and Advanced Studies (ICREA) and the FWO project G033822N.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A2': $\mathrm{PE}(u_\mathrm{d}) = L$

generically guarantee (GOAL). The question occurs: What are the nongeneric cases of A1 and A2' in which (GOAL) fails? The answered is given in Section 3.

2. RELAXATION OF THE FUNDAMENTAL LEMMA

Consider a controllable LTI systems $\mathscr{B}$ with an input/output partitioning of the variables $w = \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$. In view of A2', $\mathrm{PE}(u_\mathrm{d}) = L + k$. Next, we show that the minimal k, for which (GOAL) holds, is the controllability index ℓ_ctr of $\mathscr{B}$. Let $h(0), h(1), \ldots$ be the Markov parameters of $\mathscr{B}$ and σ, $(\sigma w)(t) := w(t+1)$ be the unit shift operator.

Lemma 1. For $w_\mathrm{d} \in \mathscr{B}|_T$ and natural numbers L, k such that $L + k \leq T$ and $k \geq \ell_\mathrm{ctr}$,

$$\begin{bmatrix} \mathscr{H}_L(\sigma^k u_\mathrm{d}) \\ \mathscr{H}_L(\sigma^k y_\mathrm{d}) \end{bmatrix} = \underbrace{\begin{bmatrix} 0_{mL \times mk} & I_{mL} \\ \mathscr{T}_\mathrm{ini} & \mathscr{T} \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} \mathscr{H}_k(u_\mathrm{d}) \\ \mathscr{H}_L(\sigma^k u_\mathrm{d}) \end{bmatrix}}_{\mathscr{H}_{k+L}(u_\mathrm{d})}, \tag{3}$$

where

$$\mathscr{T}_\mathrm{ini} := \begin{bmatrix} h(k) & h(k-1) & \cdots & h(1) \\ h(k+1) & h(k) & \cdots & h(2) \\ \vdots & \vdots & & \vdots \\ h(k+L-1) & h(k+L-2) & \cdots & h(L) \end{bmatrix} \in \mathbb{R}^{pL \times mk}$$

and

$$\mathscr{T} := \begin{bmatrix} h(0) & & & \\ h(1) & h(0) & & \\ \vdots & \ddots & \ddots & \\ h(L-1) & \cdots & h(1) & h(0) \end{bmatrix} \in \mathbb{R}^{pL \times mL}.$$

In (3), the first $k \geq \ell_\mathrm{ctrb}$ input samples $\big(u(1), \ldots, u(k)\big)$ play the role of the initial condition $x(k)$ for $\sigma^k w_\mathrm{d}$. Due to controllability any $x(k) \in \mathbb{R}^n$ can be reached by the an input $\big(u(1), \ldots, u(k)\big)$.

Theorem 1. With $L \geq \ell$, assumptions A1 and

A2": $\mathrm{PE}(u_\mathrm{d}) = L + \ell_\mathrm{ctrb}$

are necessary and sufficient for (GOAL).

3. CHARACTERIZATION OF THE NONGENERIC CASES

Lemma 2. The following are equivalent:

(1) $u_\mathrm{d} \in (\mathbb{R}^m)^T$ is persistently exciting of order $\mathrm{PE}(u_\mathrm{d}) = \ell_u$,
(2) u_d is a response of an autonomous system LTI system $\mathscr{B}_u$ with $T \geq (m+1)\ell_u - 1$ samples, *i.e.*, $u_\mathrm{d} \in \mathscr{B}_u|_T$, and for a minimal state-space representation

$$\mathscr{B}_\mathrm{ss}(A_u, C_u) := \{\, u \mid \text{there is } x_u \in (\mathbb{R}^n)^\mathbb{N}, \text{ such that } \sigma x_u = A_u x_u,\ u = C_u x_u \,\}.$$

of $\mathscr{B}_u$ with initial condition $x_{u,\mathrm{ini}} = x_u(1)$ that generates u_d, the pair $\big(A_u, x_{u,\mathrm{ini}}\big)$ is controllable.

By Lemma 2 a system $\mathscr{B}$ with m inputs, lag ℓ, and order n, which input u is persistently exciting of order ℓ_u, can be augmented with a model $\mathscr{B}_u$ of the input, resulting in an extended autonomous system $\mathscr{B}_\mathrm{ext}$ for $w = \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$. Let $\mathscr{B}_\mathrm{ss}(A_u, C_u)$ be a minimal state-space representation of the input model $\mathscr{B}_u$ and $\mathscr{B}_\mathrm{ss}(A, B, C, D)$ be a minimal input/state/output representation of $\mathscr{B}$. The extended system $\mathscr{B}_\mathrm{ext}$ is given by $\mathscr{B}_\mathrm{ext} = \mathscr{B}_\mathrm{ss}(A_\mathrm{ext}, C_\mathrm{ext})$, where

$$A_\mathrm{ext} = \begin{bmatrix} A_u & 0 \\ BC_u & A \end{bmatrix} \text{ and } C_\mathrm{ext} = \begin{bmatrix} C_u & 0 \\ DC_u & C \end{bmatrix}.$$

The extended state is $x_\mathrm{ext} = \left[\begin{smallmatrix} x_u \\ x \end{smallmatrix}\right]$, where x_u is the state of $\mathscr{B}_\mathrm{ss}(A_u, C_u)$ and x is the state of $\mathscr{B}_\mathrm{ss}(A, B, C, D)$. By using the input model $\mathscr{B}_u$, we transform the original problem about $\mathscr{B}$ into an equivalent problem about $\mathscr{B}_\mathrm{ext}$.

The following proposition derives a state transformation that block-diagonalizes A_ext. The block-diagonalization leads to a representation of $\mathscr{B}_\mathrm{ext}$ with decoupled states of $\mathscr{B}_u$ and $\mathscr{B}$. The decoupling simplifies the subsequent analysis.

Proposition 1. Assume that A and A_u have no common eigenvalues and let $V \in \mathbb{R}^{n \times n_u}$ be the solution to the equation

$$AV - VA_u = BC_u.$$

Then, $\mathscr{B}_\mathrm{ext} = \mathscr{B}_\mathrm{ss}\,(A'_\mathrm{ext}, C'_\mathrm{ext})$, where

$$A'_\mathrm{ext} = \begin{bmatrix} A_u & 0 \\ 0 & A \end{bmatrix} \text{ and } C'_\mathrm{ext} = \begin{bmatrix} C_u & 0 \\ C' & C \end{bmatrix}, \text{ with } C' := DC_u - CV.$$

The state of $\mathscr{B}_\mathrm{ss}\,(A'_\mathrm{ext}, C'_\mathrm{ext})$ is $x'_\mathrm{ext} = \left[\begin{smallmatrix} x_u \\ Vx_u + x \end{smallmatrix}\right]$, where x_u is the state of $\mathscr{B}_\mathrm{ss}(A_u, C_u)$ and x is the state of $\mathscr{B}_\mathrm{ss}(A, B, C, D)$.

Proposition 1 shows that the nongeneric cases of A1 and A2', in which (GOAL) fails correspond to a special choice of the initial condition of $\mathscr{B}$:

$$x_\mathrm{ini} = -V x_{u,\mathrm{ini}}. \tag{4}$$

By choosing the initial condition (4), we have that

$$w_\mathrm{d} = \begin{bmatrix} C_u \\ C' \end{bmatrix} \exp_{A_u} x_{u,\mathrm{ini}},$$

where $\exp_\lambda$ is the exponential function $\exp_\lambda(t) := e^{\lambda t}$. Then, $\mathrm{rank}\, \mathscr{H}_L(w_\mathrm{d}) \leq n_u$. In a trajectory w_d corresponding to (4) the transient is removed, *i.e.*, y_d, has no terms $\exp_\lambda$ where λ is an eigenvalue of A.

The characterization of the persistently exciting inputs as an output of an autonomous LTI system together with Theorem 1 allow us to do experiment design under additional constraints, such as smoothness and bounds of the input. This direction for future work is particularly useful for data-driven control of power electronics systems and will be presented elsewhere.

REFERENCES

De Persis, C. and Tesi, P. (2020). Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Trans. Automat. Contr.*, 65, 909–924.

Markovsky, I. and Dörfler, F. (2020). Identifiability in the behavioral setting. Under review.

Markovsky, I. and Dörfler, F. (2021). Behavioral systems theory in data-driven analysis, signal processing, and control. *Annual Reviews in Control*, 52, 42–64.

Markovsky, I. and Rapisarda, P. (2008). Data-driven simulation and control. *Int. J. Control*, 81(12), 1946–1959.

van Waarde, H., Eising, J., Trentelman, H., and Camlibel, K. (2020). Data informativity: A new perspective on data-driven analysis and control. *IEEE Trans. Automat. Contr.*, 65(11), 4753–4768.

van Waarde, H., et al. (2020). Willems' fundamental lemma for state-space systems and its extension to multiple datasets. *IEEE Control Systems Letters*, 4, 602–607.

van Waarde, H. (2021). Beyond persistent excitation: Online experiment design for data-driven modeling and control. *IEEE Control. Syst. Lett.*, 6, 319–324.

Willems, J.C., et al. (2005). A note on persistency of excitation. *Control Lett.*, 54(4), 325–329.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Discrete-time Hands-off Feedback Control with Real-time Optimization ⋆

Masaaki Nagahara * **Moritz Schulze Darup** **
Daniel Quevedo ***

* *The University of Kitakyushu, Hibikino 1-1, Wakamatsu, Kitakyushu, Fukuoka, 808-0135, Japan (e-mail: nagahara@ieee.org)*
** *TU Dortmund University, Leonhard-Euler-Str. 2, 44221 Dortmund, Germany (e-mail: moritz.schulzedarup@tu-dortmund.de)*
*** *Queensland University of Technology, 2 George St, Brisbane City QLD 4000, Australia (e-mail: dquevedo@ieee.org)*

Abstract: In this paper, we discuss hands-off feedback control of discrete-time linear time-invariant systems based on receding horizon control. Hands-off control, also known as sparse control, is a control that has a long time duration over which the control action is exactly zero whilst satisfying control objectives. To obtain the maximum hands-off control, the ℓ^1-norm optimization is adopted. For a model predictive control formulation, we need to numerically solve the ℓ^1 optimization with equality/inequality constraints. Although fast iterative algorithms are known to solve the optimization problem, they will often not be fast enough for control systems that need real-time computation. To obtain the control values in real time, we propose to stop the iteration for the ℓ^1 optimization after just one step. We prove that this strategy leads to practical stability of the closed-loop, provided the systems are open-loop stable. Simulation results show the effectiveness of the proposed method.

Keywords: Optimal control, sparse control, model predictive control, real-time computation, convex optimization, stability.

1. INTRODUCTION

Due to the public interest in climate change, it is crucial to consider the impact of industrial products on the environment, rather than pursuing maximum efficiency. *Maximum hands-off control* is a recently developed mathematical framework for achieving such a requirement in product design of control systems. More precisely, the control problem is to find a feasible control that has the minimum time duration in which the actuators are active. In other words, maximum hands-off control is the sparsest (or L^0-optimal) control, with which the actuators can be stopped over the inactive time duration; this results in decreased energy consumption, elimination of CO2 emissions, and reduction of noise and vibration. Accordingly, hands-off control is often called *green control*.

Maximum hands-off control has been proposed in (Nagahara et al., 2016a) for continuous-time systems, and (Nagahara et al., 2016b) for discrete-time systems. In both cases, maximum hands-off control is first designed over a finite horizon, and then implemented as self-triggered control in (Nagahara et al., 2016a), and model predictive control (MPC) in (Nagahara et al., 2016b) to realize feedback control. Stability theorems are obtained for these feedback control methods, assuming that the L^1 norm (as a surrogate of L^0 norm) or the ℓ^1 norm (as a surrogate of ℓ^0 norm) optimization can be computed immediately and exactly. For a simple plant as a double integrator, the L^1 optimal control, or minimum fuel control (Athans and Falb, 2007), is obtained in a closed form, and hence the control is almost ideally implemented.

However, if the plant is not that simple, then we need to numerically compute the optimal control by optimization algorithms, such as proximal gradient algorithm (Beck and Teboulle, 2010), or alternating direction method of multipliers (ADMM) (Boyd et al., 2011). These methods require infinitely many iterations to obtain the exact solution, and in practical applications, we need to stop the iteration in a small number to obtain a control in real time. Then, the stability should be taken into account under non-exact optimal control.

In (Parys and Pipeleers, 2018; Parys et al., 2019; Schulze Darup et al., 2019), a real-time implementation of the proximal gradient method and the ADMM in MPC has been proposed. They consider just one-step iteration for the optimization, and derive stability results under such non-exact control. This method however assumes a quadratic cost function, which is not suitable for our maximum hands-off control using the non-smooth ℓ^1 norm in the cost function. In this paper, we give stability results under one-step iteration in the algorithm for maximum hands-off feedback control using MPC, assuming that the plant is controlled stable. This assumption is not so re-

⋆ This work was partly supported by JSPS KAKENHI Grant Numbers JP20H02172 and JP20K21008. *The paper is a resubmission of our extended abstract accepted for presentation at the MTNS 2020 in Cambridge, which was later extended to the journal paper (Schulze Darup et al., 2021).*

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

strictive since stabilization is often achieved by a local controller.

Notation

We will use the following notation throughout this paper: $\mathbb{R}$ denotes the set of real numbers. For positive integers n and m, $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ denote the sets of n-dimensional real vectors and $m \times n$ real matrices, respectively. We use boldface lowercase letters, e.g. $\boldsymbol{v}$, to represent vectors, and upper case letters, e.g. A for matrices. For a positive integer n, $\mathbf{0}_n$ denotes the n-dimensional zero vector, that is, $\mathbf{0}_n = [0,\dots,0]^\top \in \mathbb{R}^n$. If the dimension is clear, the zero vector is simply denoted by $\mathbf{0}$. The superscript $(\cdot)^\top$ means the transpose of a vector or a matrix. For a matrix Φ, $\ker(\Phi)$ denotes the kernel (or the null space) of Φ. For a vector $\boldsymbol{v} = [v_1, v_2, \dots, v_n]^\top \in \mathbb{R}^n$, we define the ℓ^1 and ℓ^2 norms respectively by

$$\|\boldsymbol{v}\|_1 \triangleq \sum_{k=1}^{n} |v_k| \quad \text{and} \quad \|\boldsymbol{v}\|_2 \triangleq \sqrt{\sum_{k=1}^{n} |v_k|^2}.$$

2. FINITE-HORIZON MAXIMUM HANDS-OFF CONTROL

Here we consider a discrete-time linear time-invariant system modeled by

$$\boldsymbol{x}[k+1] = A\boldsymbol{x}[k] + \boldsymbol{b}u[k],\ k = 0, 1, 2, \dots, \tag{1}$$

where $\boldsymbol{x}[k] \in \mathbb{R}^n$ is the state, $u[k] \in \mathbb{R}$ is the scalar control input, and $A \in \mathbb{R}^{n\times n}$, $\boldsymbol{b} \in \mathbb{R}^n$.

For this system, we consider the reachability problem. That is, we find a control sequence (or vector)

$$\boldsymbol{u} \triangleq [u[0]\ u[1]\ \dots\ u[N-1]]^\top \tag{2}$$

that drives the state $\boldsymbol{x}[k]$ from a given initial state $\boldsymbol{x}[0] = \boldsymbol{x}_0$ to the origin $\boldsymbol{x}[N] = \mathbf{0}$ in N steps.

Assumption 1. We assume the following:

(1) The pair $(A, \boldsymbol{b})$ is reachable.
(2) The horizon length N is greater than or equal to n.

Under Assumption 1, there is at least one solution to the reachability problem. Actually there are infinitely many solutions if $N > n$. We call the solutions *feasible controls*. The set of feasible controls is described by

$$\mathcal{U}_{\boldsymbol{x}_0} = \left\{\boldsymbol{u} \in \mathbb{R}^N : A^N\boldsymbol{x}_0 + \Phi\boldsymbol{u} = \mathbf{0}\right\}. \tag{3}$$

where

$$\Phi \triangleq \left[A^{N-1}\boldsymbol{b}\ A^{N-2}\boldsymbol{b}\ \dots\ A\boldsymbol{b}\ \boldsymbol{b}\right]. \tag{4}$$

Among the feasible controls in $\mathcal{U}_{\boldsymbol{x}_0}$, we seek the minimum ℓ^1-norm control:

$$\underset{\boldsymbol{u}\in\mathbb{R}^N}{\text{minimize}}\ \|\boldsymbol{u}\|_1 \text{ subject to } \boldsymbol{u} \in \mathcal{U}_{\boldsymbol{x}_0}. \tag{5}$$

The solution of this ℓ^1 optimization is called the *maximum hands-off control*.

The optimization problem in (5) is effectively solved by using the alternating direction method of multipliers (ADMM) (Boyd et al., 2011). The algorithm is described as follows (Nagahara et al., 2016b):

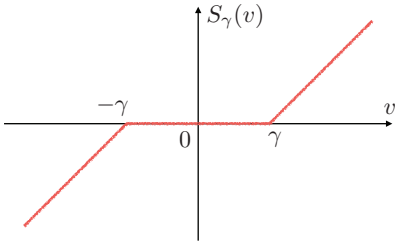


Fig. 1. Soft-thresholding operator $S_\gamma(v)$

$$\begin{aligned}\boldsymbol{v}[j+1] &:= \Pi_{\boldsymbol{x}_0}(\boldsymbol{u}[j] - \boldsymbol{w}[j]),\\ \boldsymbol{u}[j+1] &:= S_\gamma(\boldsymbol{v}[j+1] + \boldsymbol{w}[j]),\\ \boldsymbol{w}[j+1] &:= \boldsymbol{w}[j] + \boldsymbol{v}[j+1] - \boldsymbol{u}[j+1],\\ & j = 0, 1, 2, \dots,\end{aligned} \tag{6}$$

where $\gamma > 0$ is a step-size parameter, $\Pi_{\boldsymbol{x}_0}$ is the projection operator onto $\mathcal{U}_{\boldsymbol{x}_0}$, that is,

$$\Pi_{\boldsymbol{x}_0}(\boldsymbol{v}) = G\boldsymbol{v} + L\boldsymbol{x}_0, \tag{7}$$

with

$$G \triangleq I - \Phi^\top(\Phi\Phi^\top)^{-1}\Phi, \quad L \triangleq -\Phi^\top(\Phi\Phi^\top)^{-1}A^N, \tag{8}$$

and S_γ is the element-wise soft thresholding operator defined by

$$S_\gamma(v) \triangleq \begin{cases} v - \gamma, & \text{if } v > \gamma,\\ 0, & \text{if } |v| \le \gamma,\\ v + \gamma, & \text{if } v < -\gamma, \end{cases} \tag{9}$$

for scalar v. Figure 1 shows the soft-thresholding operator.

Intuitively speaking, the algorithm in (6) is a combination of the projection $\Pi_{\boldsymbol{x}_0}$ onto the feasible control set $\mathcal{U}_{\boldsymbol{x}_0}$ and the *sparsity-promoting* soft-thresholding function S_γ.

3. REAL-TIME FEEDBACK CONTROL BY MPC

Now we implement the finite-horizon control scheme in MPC to realize feedback control. We assume at the moment that the exact maximum hands-off control, the exact solution to (5) is obtained for a given state $\boldsymbol{x}_0$. By $\boldsymbol{C}_{\text{exact}}(\boldsymbol{x}_0)$ we denote the map from the state $\boldsymbol{x}_0 \in \mathbb{R}^n$ to the exact maximum hands-off control, that is,

$$\boldsymbol{C}_{\text{exact}}(\boldsymbol{x}_0) \triangleq \underset{\boldsymbol{u}\in\mathcal{U}_{\boldsymbol{x}_0}}{\arg\min} \|\boldsymbol{u}\|_1. \tag{10}$$

Then, the feedback control $u[k]$ for (1) by MPC is described by

$$u[k] = \Gamma\boldsymbol{C}_{\text{exact}}(\boldsymbol{x}[k]), \quad k = 0, 1, 2, \dots, \tag{11}$$

where $\Gamma \triangleq [1\ 0\ \dots\ 0] \in \mathbb{R}^{1\times N}$ is a matrix to extract the first element in $\boldsymbol{C}_{\text{exact}}(\boldsymbol{x}[k])$, see (Nagahara et al., 2016b) for details.

It is however difficult to obtain the exact $\boldsymbol{C}_{\text{exact}}(\boldsymbol{x}[k])$ in real time. The idea to obtain a real-time solution to the optimization is to stop the iteration in the ADMM algorithm (6) before convergence. In particular, we can simply use a *one-step iteration* for the control. We denote this process by $\boldsymbol{C}_1(\boldsymbol{x}_0; \boldsymbol{u}, \boldsymbol{v})$ with initial guesses $\boldsymbol{u}$ and $\boldsymbol{w}$, namely,

$$\begin{aligned} \boldsymbol{v}^+ &= \Pi_{\boldsymbol{x}_0}(\boldsymbol{u}-\boldsymbol{w}) \\ \boldsymbol{u}^+ &= S_\gamma(\boldsymbol{v}^+ + \boldsymbol{w}) \\ \boldsymbol{w}^+ &= \boldsymbol{w} + \boldsymbol{v}^+ - \boldsymbol{u}^+ \\ \boldsymbol{C}_1(\boldsymbol{x}_0;\boldsymbol{u},\boldsymbol{w}) &= \begin{bmatrix} \boldsymbol{v}^+ \\ \boldsymbol{u}^+ \\ \boldsymbol{w}^+ \end{bmatrix} \end{aligned} \tag{12}$$

Then the feedback control is obtained by

$$u[k] = \tilde{\Gamma}\boldsymbol{C}_1(\boldsymbol{x}[k];\boldsymbol{u}[k],\boldsymbol{w}[k]), \quad k=0,1,2,\dots \tag{13}$$

where $\tilde{\Gamma} \triangleq [\boldsymbol{0}_{1\times N}\ \Gamma\ \boldsymbol{0}_{1\times N}]$, and $\boldsymbol{u}[k]$ and $\boldsymbol{w}[k]$ are initial guesses for the one-step iteration at time k. For the initial guesses $\boldsymbol{u}[k]$ and $\boldsymbol{w}[k]$, we adopt the *warm-start* strategy that uses the previous ones, that is, they are updated by

$$\begin{aligned} \boldsymbol{u}[k+1] &= \begin{bmatrix} u_1^+[k]\ u_2^+[k]\ \dots\ u_{N-1}^+[k]\ 0 \end{bmatrix}^\top, \\ \boldsymbol{w}[k+1] &= \begin{bmatrix} w_1^+[k]\ w_2^+[k]\ \dots\ w_{N-1}^+[k]\ 0 \end{bmatrix}^\top, \end{aligned} \tag{14}$$

where $u_i^+[k]$ and $w_i^+[k]$ are respectively the $(i+1)$-th element $(i=0,1,\dots,N-1)$ of $\boldsymbol{u}^+[k]$ and $\boldsymbol{w}^+[k]$ obtained from $\boldsymbol{C}_1(\boldsymbol{x}[k];\boldsymbol{u}[k],\boldsymbol{w}[k])$.

In summary, the feedback control system is described as follows:

- one-step optimization

$$\begin{aligned} \boldsymbol{v}^+[k] &= \Pi_{\boldsymbol{x}[k]}(\boldsymbol{u}[k]-\boldsymbol{w}[k]) \\ \boldsymbol{u}^+[k] &= S_\gamma(\boldsymbol{v}^+[k]+\boldsymbol{w}[k]) \\ \boldsymbol{w}^+[k] &= \boldsymbol{w}[k] + \boldsymbol{v}^+[k] - \boldsymbol{u}^+[k] \end{aligned} \tag{15}$$

- control signal selection

$$u[k] = \Gamma\boldsymbol{u}^+[k] \tag{16}$$

- state update

$$\begin{aligned} \boldsymbol{x}[k+1] &= A\boldsymbol{x}[k] + \boldsymbol{b}u[k] \\ \boldsymbol{u}[k+1] &= D\boldsymbol{u}^+[k], \\ \boldsymbol{w}[k+1] &= D\boldsymbol{w}^+[k], \end{aligned} \tag{17}$$

  where D is the "one-step" shift matrix defined by

$$D \triangleq \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix}. \tag{18}$$

4. STABILITY ANALYSIS

From (15)–(17), the state-space equation of the feedback system is given by

$$\boldsymbol{\xi}[k+1] = \mathcal{A}\boldsymbol{\xi}[k] + \mathcal{B}S_\gamma(K\boldsymbol{\xi}[k]), \qquad \boldsymbol{\xi}[0] = \boldsymbol{\xi}_0, \tag{19}$$

with the augmented state $\boldsymbol{\xi} \triangleq [\boldsymbol{x}^\top, \boldsymbol{u}^\top, \boldsymbol{\nu}^\top]^\top$ where $\boldsymbol{\nu}[k] \triangleq \gamma\boldsymbol{w}[k]$, and the augmented system matrices

$$\mathcal{A} \triangleq \begin{bmatrix} A & 0 & 0 \\ 0 & 0 & 0 \\ DL & DG & D(I-G) \end{bmatrix}, \quad \mathcal{B} \triangleq \begin{bmatrix} \boldsymbol{b}\Gamma \\ D \\ -D \end{bmatrix} \tag{20}$$

and the augmented "controller matrix"

$$K \triangleq [L,\ G,\ I-G], \tag{21}$$

where L and G are defined in (8).

Let us consider the set

$$\mathcal{N}_\gamma \triangleq \{\boldsymbol{\xi} \in \mathbb{R}^{n+2N} : \|K\boldsymbol{\xi}\|_\infty \le \gamma\}. \tag{22}$$

Note that it contains a neighborhood of the augmented origin $\boldsymbol{0} \in \mathbb{R}^{n+2N}$. We further note that $S_\gamma(K\boldsymbol{\xi}) = \boldsymbol{0}$ for every $\boldsymbol{\xi} \in \mathcal{N}_\gamma$. Hence, on $\mathcal{N}_\gamma$, the dynamics (19) simplifies to the linear dynamics $\boldsymbol{\xi}^+ = \mathcal{A}\boldsymbol{\xi}$. Clearly, the stability of the linear dynamics depends on the eigenvalues of the matrix $\mathcal{A}$.

Theorem 1. The matrix $\mathcal{A}$ is Schur stable, if and only if A is Schur stable.

Proof. Since $\mathcal{A}$ is a block-triangular matrix, its eigenvalues correspond to the union of the eigenvalues of the diagonal blocks A, $0 \in \mathbb{R}^{N\times N}$, and

$$D(I-G) = DX, \quad X \triangleq \Phi^\top(\Phi\Phi^\top)^{-1}\Phi. \tag{23}$$

Since $\mathcal{A}$ inherits all eigenvalues of A, we immediately see that $\mathcal{A}$ is not Schur stable whenever A is not Schur stable.

It remains to show that $\mathcal{A}$ is Schur stable whenever A is Schur stable. Clearly, this relation holds if the matrix (23) is Schur. To prove that all eigenvalues $\lambda \in \mathbb{C}$ of (23) satisfy $|\lambda| < 1$, we use a result from Wielandt (1972) on eigenvalue locations for products of two matrices. As a preparation, we rewrite the matrix X using a singular value decomposition of the form $\Phi = U\Sigma V^\top$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{N\times N}$ are unitary matrices and where

$$\Sigma \triangleq \begin{bmatrix} \sigma_1 & & 0 & 0 \dots 0 \\ & \ddots & & \vdots\ \ddots\ \vdots \\ 0 & & \sigma_n & 0 \dots 0 \end{bmatrix} \in \mathbb{R}^{n\times N} \tag{24}$$

contains the singular values $\sigma_1 \ge \dots \ge \sigma_n$. Note that we have $\sigma_n > 0$ due to $\operatorname{rank}(\Phi) = n$. Using this decomposition, we easily find

$$\begin{aligned} X &= V\Sigma^\top U^\top (U\Sigma V^\top V\Sigma^\top U^\top)^{-1} U\Sigma V^\top \\ &= V\Sigma^\top U^\top U^{-\top}(\Sigma\Sigma^\top)^{-1}U^{-1}U\Sigma V^\top \\ &= V\Sigma^\top \begin{bmatrix} \sigma_1^{-2} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-2} \end{bmatrix} \Sigma V^\top = V \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} V^\top. \end{aligned} \tag{25}$$

Hence, X has n eigenvalues 1 and $N-n$ eigenvalues 0. The statements in Wielandt (1972) build on the field of values (or the numerical range)

$$\mathcal{F}(\Psi) \triangleq \{\boldsymbol{c}^\dagger\Psi\boldsymbol{c}, \quad \boldsymbol{c} \in \mathbb{C}^\ell : \boldsymbol{c}^\dagger\boldsymbol{c} = 1\} \tag{26}$$

of matrices $\Psi \in \mathbb{C}^{\ell\times\ell}$. In this context, the identified structure (25) has two important implications. First, we have

$$\mathcal{F}(X) = \operatorname{conv}\{0,1\} = [0,1] \tag{27}$$

according to (Wielandt, 1972, p. 61) since X is normal with eigenvalues 0 and 1. Second, we can apply (Wielandt, 1972, Thm. 3) since X is symmetric and positive definite. As a consequence, we have $\lambda \in \mathcal{F}(D)\,\mathcal{F}(X)$ for all eigenvalues λ of (23). Hence, taking (27) into account, $|\lambda| < 1$ is guaranteed if $\mathcal{F}(D)$ is contained in the interior of the unit disk. Since we indeed have [1]

$$\mathcal{F}(D) = \left\{ \lambda \in \mathbb{C} \;\middle|\; |\lambda| < \cos\left(\frac{\pi}{N+1}\right) \right\} \tag{28}$$

according to (Marcus and Shure, 1979, Thm. 1), the proof is complete.

[1] Using the notation in Marcus and Shure (1979), D as in (18) can be constructed based on an "injection" with no cycles and m open circuits of length N. Hence, we find $\kappa = 0$ and $\nu = N$ according to (Marcus and Shure, 1979, p. 112). Consequently, equation (8) in Marcus and Shure (1979) applies, which leads to (28).

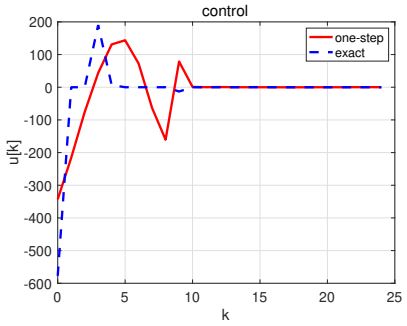


Fig. 2. Maximum hands-off control: exact control (dashed line) and control with one-step iteration (solid line).

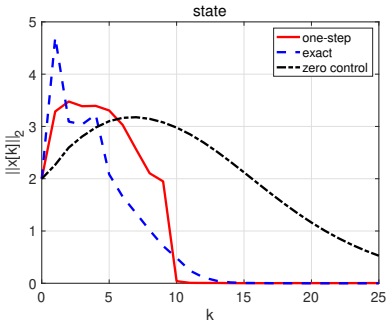


Fig. 3. 2-norm of the state $\|\boldsymbol{x}[k]\|_2$ by exact control (dashed line) and control with one-step iteration (solid line).

Remark 1. The stability result is for local dynamics around the origin. For the global dynamics, the stability is also analyzed in Schulze Darup et al. (2021).

5. NUMERICAL EXAMPLE

Let us consider a continuous-time plant

$$\dot{\boldsymbol{x}}_{\mathrm{c}}(t) = A_{\mathrm{c}}\boldsymbol{x}_{\mathrm{c}}(t) + \boldsymbol{b}_{\mathrm{c}}u_{\mathrm{c}}(t),$$

with

$$A_{\mathrm{c}} = \begin{bmatrix} -12 & -6.75 & -3.375 & -1.2656 \\ 8 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 2 & 0 \end{bmatrix}, \quad \boldsymbol{b}_{\mathrm{c}} = \begin{bmatrix} 0.125 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Note that the above state-space realization is from the transfer function $1/(s+3)^4$. Then, we discretize this plant model with sampling period $h = 0.1$ to obtain a discrete-time model as in (1) by zero-order hold discretization (or step invariant transformation (Chen and Francis, 1995)). We set the initial state $\boldsymbol{x}[0] = [1, 1, 1, 1]^\top$ and the horizon length $N = 10$.

For the one-step ADMM algorithm in (15), we set the parameter $\gamma = 1$. We also compute the exact solution $\boldsymbol{C}_{\mathrm{exact}}(\boldsymbol{x}[k])$ for comparison. This is computed by CVX (Grant and Boyd, 2008, 2014) on MATLAB. Figure 2 shows the obtained control signals.

Since A is stable, the zero control $u \equiv 0$ is also a stabilizing control. That is, with this zero control, the state converges to the origin from any initial state. Although the zero control is obviously the sparsest control, the convergence rate achieved depends only on the time constant, or the largest absolute value of the eigenvalues of A. Figure 3 shows the 2-norm of the state $\boldsymbol{x}[k]$, $k = 0, 1, 2, \ldots, N$ by the proposed one-step control, the exact control, and the zero control. We can see that they all converge to zero. We can also see that the state by the zero-control converges much slower than the hands-off control. The latter converges to zero in finite time.

6. CONCLUSION

In this paper, we have considered discrete-time hands-off control within an MPC formulation. Having real-time applications in mind, we have proposed the one-step optimization with a warm-start and derived a stability result of the combined system when the controlled plant is stable. Future work includes the stability analysis for unstable plants.

REFERENCES

Athans, M. and Falb, P.L. (2007). *Optimal Control*. Dover Publications. An unabridged republication of the work published by McGraw-Hill in 1966.

Beck, A. and Teboulle, M. (2010). Gradient-based algorithms with applications to signal-recovery problems. In *Convex optimization*. Cambridge University Press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 1–122.

Chen, T. and Francis, B.A. (1995). *Optimal Sampled-Data Control Systems*. Springer.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura (eds.), *Recent Advances in Learning and Control*, volume 371 of *Lecture Notes in Control and Information Sciences*, 95–110. Splinger.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

Marcus, M. and Shure, B.N. (1979). The numerical range of certain 0, 1-matrices. *Linear and Multilinear Algebra*, 7, 111–120.

Nagahara, M., Quevedo, D.E., and Nešić, D. (2016a). Maximum hands-off control: a paradigm of control effort minimization. *IEEE Trans. Autom. Control*, 61(3), 735–747.

Nagahara, M., Østergaard, J., and Quevedo, D.E. (2016b). Discrete-time hands-off control by sparse optimization. *EURASIP Journal on Advances in Signal Processing*, 2016(1), 1–8.

Parys, R.V. and Pipeleers, G. (2018). Real-time proximal gradient method for linear MPC. In *2018 European Control Conference (ECC)*, 1142–1147.

Parys, R.V., Verbandt, M., Swevers, J., and Pipeleers, G. (2019). Real-time proximal gradient method for embedded linear MPC. *Mechatronics*, 59, 1–9.

Schulze Darup, M., Book, G., and Giselsson, P. (2019). Towards real-time ADMM for linear MPC. In *2019 18th European Control Conference (ECC)*, 4276–4282.

Schulze Darup, M., Book, G., Quevedo, D.E., and Nagahara, M. (2021). Fast hands-off control using ADMM real-time iterations. *IEEE Trans. Autom. Control*.

Wielandt, H. (1972). On the eigenvalues of $A+B$ and AB. *Journal of Research of the National Bureau of Standards – B. Mathematical Sciences*, 77B, 61–63.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Modeling the economic effects of the Covid-19 pandemic in a data-driven agent-based framework

Anton Pichler

Complexity Science Hub, 1090 Vienna, Austria
(e-mail: pichler@csh.ac.at)

Abstract: We introduce a dynamic disequilibrium agent-based model (ABM) that was used to forecast the economics of the Covid-19 pandemic. This model was designed to understand the upstream and downstream propagation of the industry-specific demand and supply shocks caused by Covid-19, which were exceptional in their severity, suddenness and heterogeneity across industries. We used this model to forecast sectoral and aggregate economic activity for the United Kingdom during the early phase of the pandemic. This work demonstrates that an out of equilibrium model calibrated against national accounting data can serve as a useful real time policy evaluation and forecasting tool.

We further extend this modeling framework to a large-scale, data-driven ABM of the New York metropolitan area that simulates both, epidemic and economic outcomes across industries, occupations, and income levels. This coupled epidemic-economic model is designed to address the potential tradeoff between economy and health which has been a key issue faced by policymakers. Our results show that lockdown policies affect different social groups very heterogeneously in terms of income and infections.

Keywords: production function, shock propagation, production network, synthetic population, distributional effects

1. INTRODUCTION

The social distancing measures imposed to combat the first wave of the Covid-19 pandemic created severe industry-specific disruptions to economic output. Some industries were shut down almost entirely by lack of demand, labor shortages restricted others, and many were initially largely unaffected. Feedback effects then amplified the initial shocks. The lack of demand for final goods such as restaurants or transportation propagated upstream, reducing demand for the intermediate goods that supply these industries. Supply constraints due to a lack of labor under social distancing propagated downstream, creating input scarcity that sometimes limited production even in cases where the availability of labor and demand would not have been an issue. The resulting supply and demand constraints interacted to create bottlenecks in production, which in turn led to unemployment, eventually decreasing consumption and causing additional amplification of shocks that further decreased final demand. The unprecedented scale and heterogeneity of the shocks caused a major disruption of the economy that presented a challenge for economic modelers.

Here, we summarize our recent (and still ongoing) work of modeling the economics of the Covid-19 pandemic which aims at addressing these challenges. This abstract is mainly based on a series of papers (del Rio-Chanona et al., 2020, Pichler et al., 2020, Pichler & Farmer, 2021, Pichler et al., 2022 and Pangallo et al., 2022) with a focus on (1) a dynamic disequilibrium model of the UK production network and (2) a disaggregated, large-scale coupled epidemic-economic model of the New York metropolitan area.

2. FORECASTING THE PROPAGATION OF PANDEMIC SHOCKS

We introduce a dynamic input-output model that addresses the unique features of the pandemic. The model, which is directly initialized from national accounts and other data sources where this is possible, has several new elements that affect production, consumption and changes in the labor force. We developed this model during March-April 2020, and used it to forecast the economic consequences of the relaxation of the lockdown in the UK in real-time, in a working paper we released in May 2020 (Pichler et al., 2020). Here we show that the model predicted aggregate economic effects very well and analyze why it succeeded. We first analyze how the model anticipated the impact of the Covid-19 pandemic on the UK economy, particularly at the sectoral level. We then show that our model accurately captures supply chain effects that explain the dynamics of related industries.

The Covid-19 episode was exceptional: (1) Pandemic shocks were highly heterogeneous across industries, making it necessary to model the economy at the sectoral level, taking sectoral inter-dependencies into account; (2) the shocks affected both supply and demand simultaneously, and led to both upstream and downstream propagation; (3) the shocks were so strong and were imposed and relaxed so quickly that the economy never had time to converge to a new steady state, making dynamic models better suited than static models.

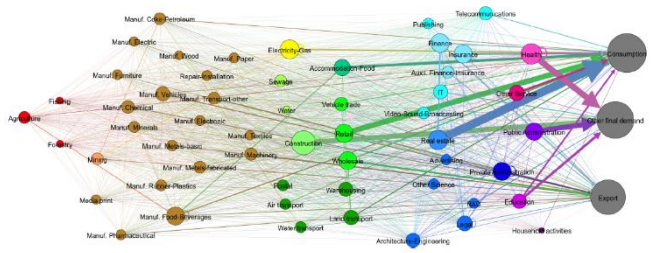


Fig. 1. The UK economy represented as a 55 sectors production network.

A key innovation in our economic model is the sector-specific treatment of the production function. We introduce a new production function that distinguishes between critical and non-critical inputs at the level of the 55 industries in the World Input-Output tables. The *partially binding Leontief* production function that we introduce here allows firms to keep producing as long as they have the inputs that are absolutely necessary. We show that a realistic specification of the production function is a key ingredient with strong effects on model accuracy.

Another key element of our modeling approach is a detailed representation of industry-specific input inventories. Inventories act as buffers in the presence of supply chain disruptions or demand shocks and thus can play an important role in shock propagation dynamics. We use a survey by the UK Office for National Statistics (ONS) on industry-level inventories to initialize our model. Our results are in line with previous work which has shown that inventory levels strongly affect the scale and dynamics of shock propagation.

We introduce a Covid-19-specific treatment of consumption. Most models do not incorporate the demand shocks that are caused by changes in consumer preferences in order to minimize risk of infection. We consider demand shocks to consumption due to "fear of infection" and consider the effect of the drop in current income due to unemployment and reduced expectations of permanent income due to pessimism about the end of the pandemic.

Finally, compared to other economic disaster models, our model explicitly considers labor. Industries adjust their labor force depending on supply constraints due to lockdown, lack of demand or lack of intermediate inputs. Adjustment is sluggish, so firms cannot instantly increase production if they lack workers, as hiring takes time.

We released our results for the UK economy online on May 21, 2020, not long after social distancing measures first began to take effect in March. Our central scenario considered a government policy for reopening that was very close to what the UK government decided. In that scenario, we predicted a 21.5% contraction of GDP in the UK economy in the second quarter of 2020 with respect to the last quarter of 2019. This forecast was remarkably close to the actual contraction of 22.1%.

In our recent work (Pichler et al. 2022) we show which of the modeling factors have been key for making good aggregate sectoral predictions, not just for GDP but also for other key economic variables.

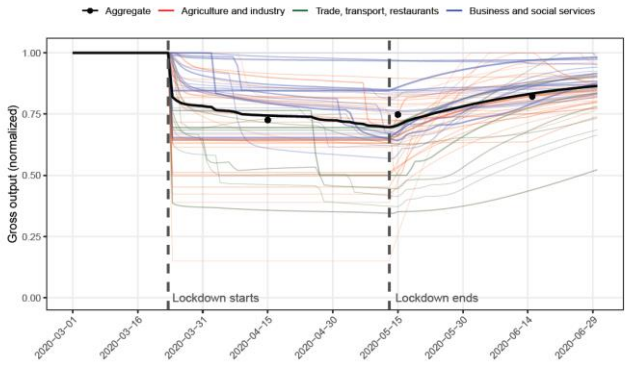


Fig. 2. Economic production as predicted by the model.

3. MODELING DISTRIBUTIONAL EFFECTS OF THE PANDEMIC

Since the outbreak of the Covid-19 pandemic, governments worldwide have successfully slowed down the transmission of the virus by enacting non-pharmaceutical interventions. These interventions include the shutdown of some customer-facing economic activities, e.g. entertainment and restaurants, and the imposition of work-from-home mandates. Such protective measures have distributional effects, i.e. heterogeneous outcomes across socio-economic groups. For instance, workers who can work from home become less likely to be infected after the imposition of these measures, while essential workers remain at risk. At the same time, these measures have different distributional economic effects depending on the industry and occupation of the workers. For example, low-income workers are more likely to work in customer-facing industries and perform in-person occupations, leading to higher risk of unemployment when these industries are closed.

Addressing the effectiveness of non-pharmaceutical interventions over behavioral change, both at the aggregate and distributional level, requires building theoretical, mechanistic models that jointly simulate epidemic and economic dynamics at a fine-grained level.

Here, we introduce an agent-based model (ABM) that simulates epidemic and economic outcomes of a large synthetic population in a metropolitan area. The socio-economic characteristics and the consumption and contact patterns are initialized from detailed census, survey, and mobility data, while the structure of the economy is initialized from input-output tables and national and regional accounts.

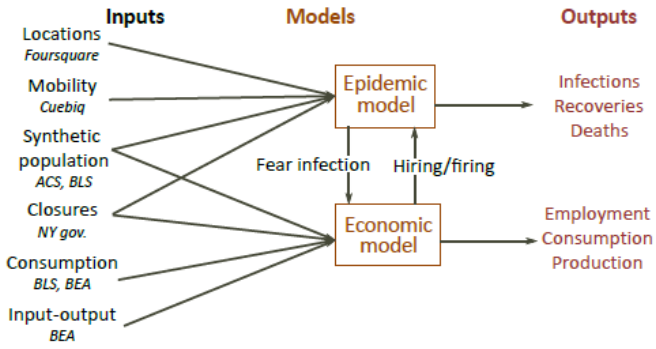


Fig. 3. Basic relationships between inputs and outputs in the epidemic-economic model.

Our joint epidemic-economic ABM merges and largely extends our former epidemic (Aleta et al. 2020) and economic (Pichler et al. 2020) models built to understand the effects of the Covid-19 pandemic and study the effectiveness of policy responses. We calibrate our model to the first wave of the pandemic in the New York metropolitan area and show that it quantitatively matches key epidemic and economic statistics, both in the aggregate and across income levels and industries. We then use our ABM to understand the interplay of behavior change and non-pharmaceutical interventions. Our key result is that strong behavior change, alike strict closure of economic activities, harms the economy but reduces infections and, thus, saves lives. This equivalence between behavior changes and closures also holds at the distributional level: We find that under strong behavior change and strict non-pharmaceutical interventions, low-income workers are more likely to become unemployed but less likely to become infected. The mechanisms behind these results suggest that a model like ours, initialized from various granular datasets, can be used to design policies that minimize health and economic damages to disadvantaged socio-economic groups, for instance by designing income support schemes that are specific to the industry and occupation of workers.

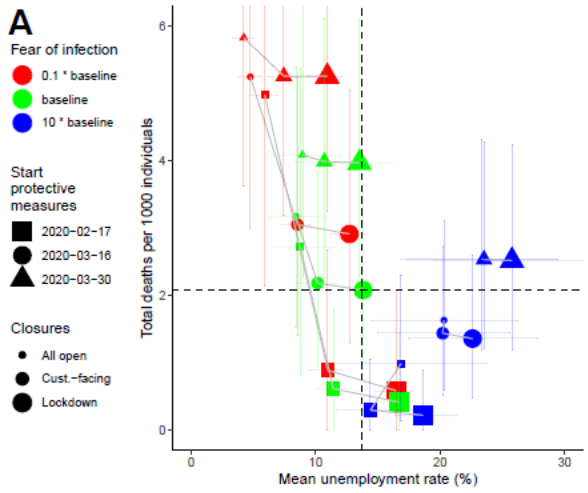


Fig. 4. Simulation results showing total deaths per 1000 individuals (y-axis) and the mean unemployment rate (x-axis) for different parametrizations.

REFERENCES

Aleta, A., Martin-Corral, D., Pastore y Piontti, A., Ajelli, M., Litvinova, M., Chinazzi, M., ... & Moreno, Y. (2020). Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature Human Behaviour*, 4(9), 964-971.

del Rio-Chanona, R. M., Mealy, P., Pichler, A., Lafond, F., & Farmer, J. D. (2020). Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. *Oxford Review of Economic Policy*, *36*, 94-137.

Pangallo, M., Aleta A., del Rio-Chanona, R. M., Pichler, A., ... & Vespignani, A., & Farmer, J.D. (2022). Modeling the distributional epidemic-economic effects of the Covid-19 pandemic. *In preparation.*

Pichler, A., Pangallo, M., del Rio-Chanona, R. M., Lafond, F., & Farmer, J. D. (2020). Production networks and epidemic spreading: How to restart the UK economy? *Covid Economics 23*, 79–151

Pichler, A., & Farmer, J. D. (2021). Simultaneous supply and demand constraints in input–output networks: The case of Covid-19 in Germany, Italy, and Spain. *Economic Systems Research*, 1-21.

Pichler, A., Pangallo, M., del Rio-Chanona, R. M., Lafond, F., & Farmer, J. D. (2022). Forecasting the propagation of pandemic shocks with a dynamic input-output model. *R&R at Journal of Economic Dynamics and Control*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

max-p optimal boundary control of gas flow

Martin Gugat * **Michael Schuster** **

* *Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department of Data Science, Lehrstuhl für Dynamics, Control and Numerics (Alexander von Humboldt-Professur), Cauerstr. 11, 91058 Erlangen, Germany (e-mail: martin.gugat@fau.de).*
** *Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department of Data Science, Lehrstuhl für Dynamics, Control and Numerics (Alexander von Humboldt-Professur), Cauerstr. 11, 91058 Erlangen, Germany (e-mail: michi.schuster@fau.de).*

Abstract: In the transition to renewable energy sources, hydrogen will potentially play an important role for energy storage. The efficient transport of this gas is possible via pipelines. An understanding of the possibilities to control the gas flow in pipelines is one of the main building blocks towards the optimal use of gas.
For the operation of gas transport networks it is important to take into account the randomness of the consumers' demand, where often information on the probability distribution is available. Hence in an efficient optimal control model the corresponding probability should be included and the optimal control should be such that the state that is generated by the optimal control satisfies given state constraints with large probability. We comment on the modelling of gas pipeline flow and the problems of optimal nodal control with random demand, where the aim of the optimization is to determine controls that generate states that satisfy given pressure bounds with large probability. We include the H^2 norm of the control as control cost, since this avoids large pressure fluctuations which are harmful in the transport of hydrogen since they can cause embrittlement of the pipeline metal.

Keywords: gas pipeline flow, nodal control, boundary control, optimal control, hyperbolic differential equation, random demand, state constraints, pressure bound, classical solutions

1. INTRODUCTION

The isothermal Euler equations (see e.g. Banda et al. (2006), Gugat and Herty (2022))

$$\begin{cases} \rho_t + q_x = 0, \\ q_t + \left(p + \dfrac{q^2}{\rho}\right)_x = -\dfrac{1}{2}\theta \dfrac{q\,|q|}{\rho} \end{cases} \tag{1}$$

are a well-established model for gas pipeline flow, where ρ denotes the gas density, p the pressure, q the mass flow rate and $\theta \geq 0$ is a friction parameter. At the end $x = 0$ the flow rate that is desired by the consumers is given by a random variable, so we have $q(t,\, 0) = q_r(\omega)$ on a probability space $(\Omega,\, \mathcal{A},\, \mathbb{P})$. Here we assume that $q_r(\omega) \in C^1([0,\, T])$ for all $\omega \in \Omega$. Due to the influence of the random boundary term, also the pde solution becomes a random variable. At the end $x = L$ of the pipe, the pressure is controlled, $p(t,\, L) = u(t)$. We consider controls $u \in H^2([0,\, T])$. For the deterministic case, in Gugat and Sokolowski (2022), a similar optimal control problem for gas networks is considered and the existence of an optimal control is shown. See Göttlich and Schillinger (2021) for a related study for linear systems.

⋆ This work was funded by the DFG, TRR 154, *Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks*, projects C03 and C05, Projektnummer 239904186

2. THE SYSTEM

Let a time horizon $T > 0$ be given. For the case of ideal gas where $p = a^2\rho$ with the sound speed $a > 0$, our system is governed by the initial boundary value problem

$$(\mathbf{S}^\omega) \begin{cases} q(0,x) = q_0(x), \;\; \rho(0,x) = \rho_0(x),\; x \in (0,\, L), \\ q(t,\, 0) = q_r(\omega), \;\; p(t,\, L) = u(t),\; t \in (0,\, T), \\ \begin{pmatrix} \rho \\ q \end{pmatrix}_t + \begin{pmatrix} 0 & 1 \\ a^2 - \dfrac{q^2}{\rho^2} & 2\,\dfrac{q}{\rho} \end{pmatrix} \begin{pmatrix} \rho \\ q \end{pmatrix}_x = \begin{pmatrix} 0 \\ -\dfrac{\theta}{2}\dfrac{q\,|q|}{\rho} \end{pmatrix}. \end{cases}$$

Let $R_0 > 0$ denote a constant reference density. Due to the theory of semi-global solutions (see Li (2010), Li et al. (2016)) for any given time horizon $T > 0$ there exist numbers $\varepsilon(T) > 0$ and $C_1(T) > 0$ such that for all $R \geq R_0$ and all initial states that satisfy

$$\max\{\|q_0\|_{C^1([0,\, L])},\; \|\rho_0 - R\|_{C^1([0,\, L])}\} \leq \varepsilon(T) \tag{2}$$

and all $q_r(\omega)$ with

$$\|q_r(\omega)\|_{C^1([0,\, T])} \leq \varepsilon(T) \tag{3}$$

and all controls with

$$\|u - a^2\, R\|_{C^1([0,\, T])} \leq \varepsilon(T) \tag{4}$$

that are C^1-compatible with the initial state $(q_0,\, \rho_0)$ there exists a classical solution $(q^\omega,\, \rho^\omega)$ of $(\mathbf{S}^\omega)$ on $[0,\, T]$ that satisfies the a priori estimate

$\max\{\|q^\omega\|_{C^1([0,\, T]\times[0,L])}, \|\rho^\omega - R\|_{C^1([0,\, T]\times[0,L])}\} \leq C_1(T)$
$\max\{\|q_0\|_{C^1([0,\, L])},\; \|\rho_0 - R\|_{C^1([0,\, L])}, \|q_r(\omega)\|_{C^1([0,\, T])},\; \|u -$

$a^2 R\|_{C^1([0,T])}\}$. Moreover, the state depends continuously in $(C^1([0,T]\times[0,L]))^2$ on the control $u \in C^1([0,T])$.

To guarantee that a regular solution exists, in the optimal control problem the control constraint (4) and the C^1-compatibility conditions with the initial state are prescribed.

3. THE OPTIMAL CONTROL PROBLEM

Let a lower pressure bound $p_{\min} > 0$ be given. For a control $u \in H^2(0,T)$ define the objective function

$$J(u,R) = \|u-R\|_{H^2(0,T)} \qquad (5)$$
$$-\ln\left(\mathbb{P}\left(\|(p_{\min}-p^\omega)_+\|_{C([0,T]\times[0,L])}=0\right)\right).$$

The optimal control problem $\mathbf{P_{dyn}}(T)$ is to minimize $J(u,R)$ subject to the constraints $R \geq R_0$, (4) and the C^1-compatibility conditions for u, where (p_ω, q_ω) solves $(\mathbf{S}^\omega)$.

The H^2-term in the objective function helps to avoid large pressure fluctuations in the pipe that can be harmful if the gas contains hydrogen due to the danger of embrittlement, see Guy et al. (2021). The optimal control of gas transportation systems is a classical topic in process engineering, see for example Osiadacz and Swierczewski (1994).

3.1 Existence of solutions

Theorem 1. Let $(q_0,\rho_0) \in \left(C^1([0,T])\right)^2$ be given such that $a^2\rho_0 > p_{\min}$ and (2) holds. Assume that (3) and the C^1-compatibility conditions of $q_r(\omega)$ and the initial data hold almost surely. Then an optimal control that solves $\mathbf{P_{dyn}}(T)$ does exist in $(0,\infty)\times H^2([0,T])$.

Proof. The set of admissible controls is non-empty, since for all $R > 0$ there exists a control $\hat{u} \in H^2(0,T)$ that is compatible with (q_0,ρ_0) and satisfies (4). The a priori estimate implies that if R is sufficiently large, we have almost surely $p^\omega = a^2R + a^2(\rho^\omega - R) \geq a^2R - a^2C_1(T)\,\varepsilon(T) \geq p_{\min}$. Hence if R is sufficiently large, there exists a control where the objective function attains a finite value. (Note that this does not require that $p^\omega \geq p_{\min}$ almost surely, but only that $p^\omega \geq p_{\min}$ has a nonzero probability.)

The H^2-norm is a weakly sequentially lower semi-continuous functional in $H^2(0,T)$. Results from Farshbaf-Shaker et al. (2018) imply that the probabilistic part of the objective function is also weakly sequentially lower semi-continuous in $H^2(0,T)$. This can be seen as follows. A sequence that converges weakly in $H^2(0,T)$ converges strongly in $C^1([0,T])$ to a limit point $u^* \in H^2(0,T)$. Due to the theory of semi-global solutions, (2), (3) and (4) imply that the controls generate classical solutions of $(\mathbf{S}^\omega)$ almost surely and the strong convergence in $C^1([0,T])$ of the controls implies that also the corresponding subsequence of generated states given by the classical solutions of $(\mathbf{S}^\omega)$ converges strongly in $\left(C^1([0,T]\times[0,L])\right)^2$ to the solution that is generated by the limit point u^*. Then Lemma 2 in Farshbaf-Shaker et al. (2018) implies that the probability is weakly sequentially upper semi-continuous in $H^2(0,T)$. This implies that the objective functional is a weakly sequentially lower semi-continuous functional in $(0,\infty)\times H^2(0,T)$.

We consider a minimizing sequence of feasible controls (R_k, u_k). Due to the H^2-term and (4), this sequence is bounded in $\mathbb{R}\times H^2(0,T)$. Hence it contains a subsequence that converges weakly in $\mathbb{R}\times H^2(0,T)$ and thus converges strongly in $\mathbb{R}\times C^1([0,T])$ to a limit point $(R^*, u^*) \in \mathbb{R}\times H^2(0,T)$. Moreover, this also implies that u^* satisfies (4) and that the values of the objective function $J(u^*)$ is minimal.

3.2 Numerical approaches

For the numerical solution, a kernel density estimator should be used to obtain a differentiable approximation of the objective function similar to the approach in Schuster et al. (2021). The controls are represented as Fourier series, $u(t) = \frac{a_0}{2} + \sum_{j=1}^{\infty} a_j \cos\left(j\frac{2\pi}{T}t\right) + b_j \sin\left(j\frac{2\pi}{T}t\right)$. Then we have $\frac{2}{T}\|u\|^2_{H^2(0,T)}$ $= \frac{a_0^2}{2} + \sum_{j=1}^{\infty}\left(1+\left(j\frac{2\pi}{T}\right)^2+\left(j\frac{2\pi}{T}\right)^4\right)\left(|a_j|^2+|b_j|^2\right)\Big)$. Truncation of the Fourier series after a finite number of modes leads to a semi-infinite optimization problem, for a survey see Stein (2012).

4. A NUMERICAL EXAMPLE

Gas network optimization has been of interest for decades, see e.g., Herty and Sachers (2007); Zlotnik et al. (2015) for a semilinear hyperbolic gas transport model and Mak et al. (2019) for a parabolic gas transport model. But gas network optimization with H^2 control and probabilistic terms in the objective function was not considered yet. We present a numerical example on a single edge for both, a probabilistic objective function with a L^2 control term and a probabilistic objective function with an H^2 control term.

We consider the isothermal Euler equations for ideal gases, i.e., for $(t,x) \in [0,T]\times[0,L]$ we have

$$p(t,x) = a^2\rho(t,x),$$

where a denotes the speed of sound in the gas. Due to the proportionality of pressure and density we consider density control at $x = L$ instead of pressure control. All values and constants are given in *Table 1.*

Letter	Value	Unit
T	12	h
L	30	km
a	343	m/s
θ	0.2	
R	46.3	kg/m^3
$\rho_{\min}$	40.4	kg/m^3

Table 1. Values for the numerical example.

At the end $x = 0$ we assume random gas outflow. Therefor we define a deterministic function

$$q_D(t) = -\frac{16}{\pi}\sin\left(\frac{\pi}{12\cdot 60^2}t\right)\cdot\frac{16}{\pi}\cos\left(\frac{\pi}{8\cdot 60^2}t\right) + 140.$$

Let

$$\xi \sim \mathcal{N}\left(1, \sqrt{0.1}\right),$$

be a Gaussian distributed random variable on an appropriate probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Similar as in Schuster et al. (2021) we write $q_D(t)$ as Fourier series and multiply every Fourier term with a random number $\xi(\omega)$, $\omega \in \Omega$. For the implementation we cut the Fourier series after 10 terms. A sample of 20 random boundary functions $q_r(\omega)$ and the corresponding deterministic function $q_D(t)$ are shown in *Figure 1.* For the implementation we use the negative flow values since gas is transported from the end of the pipe $(x = L)$ to its beginning $(x = 0)$.

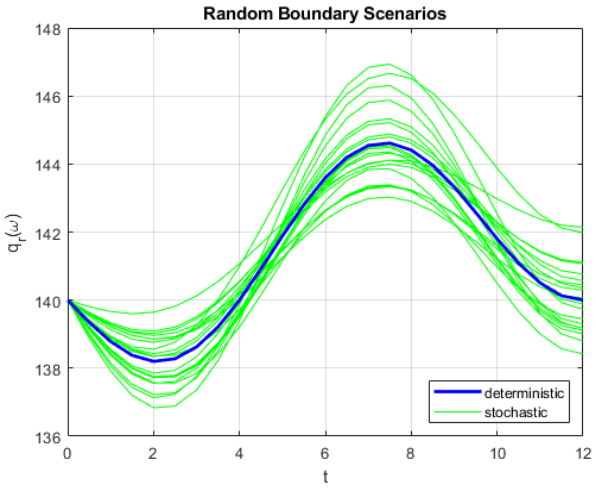


Fig. 1. Sample of 20 random boundary functions $q_r(\omega)$.

For the initial state we solve the stationary isothermal Euler equations

$$\begin{cases} q_x^\sigma = 0, \\ \left(a^2\rho^\sigma + \dfrac{(q^\sigma)^2}{\rho^\sigma}\right)_x = -\dfrac{1}{2}\theta\dfrac{q^\sigma\,|q^\sigma|}{\rho^\sigma}, \end{cases} \tag{6}$$

with the boundary conditions $q^\sigma(0) = q_D(0)$ and $\rho^\sigma(L) = 46.3\ kg/m^3$. The solution $(\rho^\sigma_{\text{init}}, q^\sigma_{\text{init}})$ of (6) serves as initial state for the dynamic problem.

The probabilistic term in the objective function is computed with a kernel density estimator approach (see Schuster et al. (2021)). Due to the friction along the pipe and due to the choice of initial states, for every time the density is minimal at $x = 0$. Thus we have $\rho^\omega(t,x) \geq \rho_{\min}$ iff $\rho^\omega(t,0) \geq \rho_{\min}$. We discretize the time interval using n_T+1 equidistant points $0 = t_0 < \cdots < t_{n_T} = T$ and we use a multivariate kernel density estimator approach with Gaussian product kernels to approximate the probabilistic term. For

$$\mathcal{P}_{\min} := \otimes_{i=1}^{n_T}[\rho_{\min}, \infty),$$

we have

$$\mathbb{P}\big(\ \rho^\omega(t,0) \geq \rho_{\min} \quad \forall t \in [0,T]\ \big) \approx$$
$$\int_{\mathcal{P}_{\min}} \frac{1}{N\sqrt{\det H}} \sum_{i=1}^{N}\prod_{j=1}^{n_T}\frac{1}{\sqrt{2\pi}} \exp\left(-\left(\frac{z_j - \rho_i(t_j)}{\sqrt{2\ H_{j,j}}}\right)^2\right)$$
$$= \frac{1}{N\ 2^{n_T}}\sum_{i=1}^{N}\prod_{j=1}^{n_T}\left[1 - \operatorname{erf}\left(\frac{\rho_{\min} - \rho_i(t_j)}{\sqrt{2H_{j,j}}}\right)\right].$$

Here N is the number of samples, $\rho_i(t_j)$ is the density for the i-th sample at $(t,x) = (t_j, 0)$ and H is a diagonal positive definite bandwidth matrix.

We define objective functions

$$J_{L^2}(u,R) = w_1\|u - R\|_{L^2(0,T)} - \ln\Big(\mathbb{P}\big(\ \rho^\omega(t,0) \geq \rho_{\min} \quad \forall t \in [0,T]\ \big)\Big),$$

and

$$J_{H^2}(u,R) = w_1\|u - R\|_{L^2(0,T)} + w_2\|u'\|_{L^2(0,T)} + w_3\|u''\|_{L^2(0,T)} - \ln\Big(\mathbb{P}\big(\ \rho^\omega(t,0) \geq \rho_{\min} \quad \forall t \in [0,T]\ \big)\Big),$$

with weights

$$w_1 = 2\cdot 10^{-3}, \quad w_2 = 1\cdot 10^{5}, \quad w_3 = 1\cdot 10^{12}.$$

The optimal controls for both objective functions ($N = 20$ and $n_T = 25$) are shown in *Figure 2.* The blue line shows the optimal density control for $J_{L^2}(u,R)$ and the red line in shows the optimal density control for $J_{H^2}(u,R)$. The results can be interpreted as follows: For an objective function without probabilistic term the optimal solution would obviously be $u \equiv R$. The density at $x = 0$ is only lower than $\rho_{\min}$ for the peak around 7 hours (cf. *Figure 1*). Thus the control only needs to be active in this time span. As it was expected the H^2 control is smoother than the L^2 control.

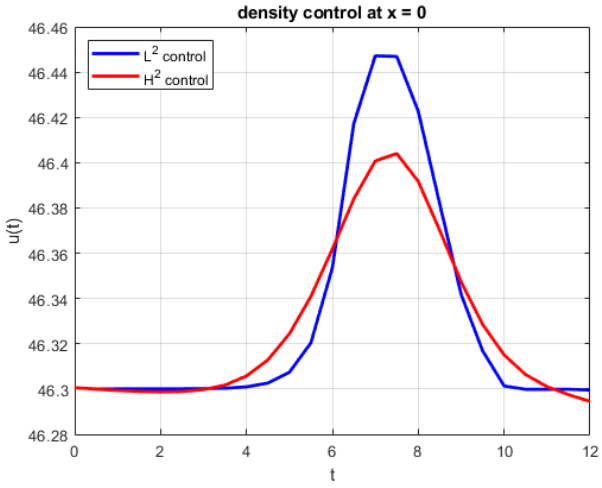


Fig. 2. Optimal control for $J_{L^2}(u,R)$ and $J_{H^2}(u,R)$.

We have

$$\mathbb{P}_{L^2}\big(\ \rho^\omega(t,0) \geq \rho_{\min} \quad \forall t \in [0,T]\ \big) \approx 74\%,$$

and

$$\mathbb{P}_{H^2}\big(\ \rho^\omega(t,0) \geq \rho_{\min} \quad \forall t \in [0,T]\ \big) \approx 66\%.$$

Thus a slight decrease of the probability leads to a smoother density control and less density fluctuations. This can also be seen in *Figure 3* and *Figure 4*. The peaks around 7 hours are smoother in *Figure 4* than in *Figure 3*. The blue line shows $\rho_{\min}$. The optimal density control and the corresponding densities at $x = 0$ would be even smoother if we would increase the weights w_2, w_3 for the L^2-Norm of the control derivatives.

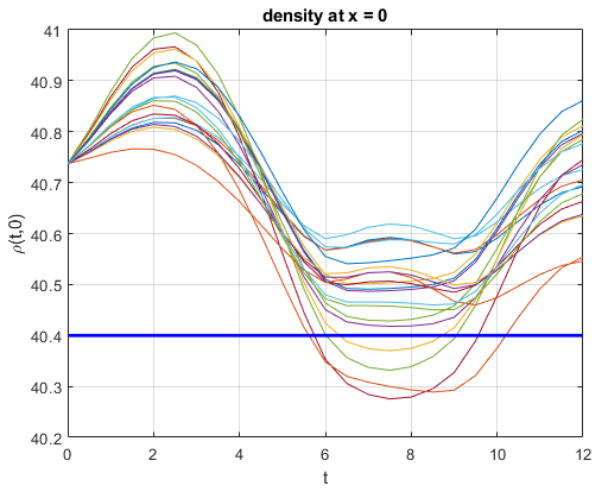


Fig. 3. Scenarios at $x = 0$ for the optimal density control of $J_{L^2}(u, R)$

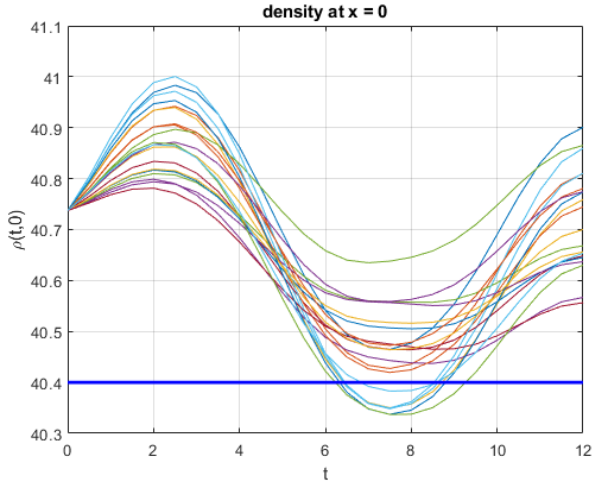


Fig. 4. Scenarios at $x = 0$ for the optimal density control of $J_{H^2}(u, R)$

5. CONCLUSION

In optimal control problems, it is important to take into account the uncertainty of the problem data in order to obtain controls that work sufficiently well in the set of data that is expected. Since in many applications information on the probability distribution of the data is available, this information should be used in an optimal control model. In our contribution we choose the probability that state constraints are satisfied as a part of the objective function. In this way, it is ensured that the optimization generates controls that are robust in the sense that the pressure bounds are satisfied with a high probability. We include an H^2 control cost in the objective functional, which is of particular interest in the context of hydrogen transport. It also serves as a Tychonov regularization term that is important for the proof of the existence of optimal controls.

ACKNOWLEDGEMENTS

This work was supported by Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Centre CRC/Transregio 154, Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks, Projects C03 and C05, Projektnummer 239904186.

REFERENCES

Banda, M.K., Herty, M., and Klar, A. (2006). Coupling conditions for gas networks governed by the isothermal Euler equations. *Netw. Heterog. Media*, 1(2), 295–314.

Farshbaf-Shaker, H., M.H., Henrion, R., and Hömberg, D. (2018). Properties of chance constraints in infinite dimensions with an application to pde constrained optimization. *Set-Valued Var. Anal*, 26, 821–841.

Gugat, M. and Sokolowski, J. (2022). On problems of dynamic optimal nodal control for gas networks. *Pure and Applied Functional Analysis*, x, x–x.

Gugat, M. and Herty, M. (2022). Modeling, control, and numerics of gas networks. In E. Trelat and E. Zuazua (eds.), *Handbook of Numerical Analysis*. doi: 10.1016/bs.hna.2021.12.002.

Guy, P., Laszlo, T., and Julien, C. (2021). Effects of hydrogen addition on design, maintenance and surveillance of gas networks. *Processes*, 9(7). doi:10.3390/pr9071219.

Göttlich, S. and Schillinger, T. (2021). Control strategies for transport networks under demand uncertainty. ArXiv preprint arXiv:2111.09674.

Herty, M. and Sachers, V. (2007). Adjoint calculus for optimization of gas networks. *Networks and Heterogeneous Media*, 2(4), 733–750.

Li, T. (2010). *Controllability and observability for quasilinear hyperbolic systems*, volume 3 of *AIMS Series on Applied Mathematics*. American Institute of Mathematical Sciences (AIMS), Springfield, MO; Higher Education Press, Beijing.

Li, T., Wang, K., and Gu, Q. (2016). *Exact boundary controllability of nodal profile for quasilinear hyperbolic systems*. SpringerBriefs in Mathematics. Springer, Singapore. doi:10.1007/978-981-10-2842-7.

Mak, T., Hentenryck, P.V., Zlotnik, A., and Bent, R. (2019). Dynamic compressor optimization in natural gas pipeline systems. *Informs Journal on Computing*, 31, 40–65.

Osiadacz and Swierczewski (1994). Optimal control of gas transportation systems. In *1994 Proceedings of IEEE International Conference on Control and Applications*, 795–796 vol.2. doi:10.1109/CCA.1994.381219.

Schuster, M., Strauch, E., Gugat, M., and Lang, J. (2021). Probabilistic constrained optimization on flow networks. *Optimization and Engineering*. doi:10.1007/s11081-021-09619-x.

Stein, O. (2012). How to solve a semi-infinite optimization problem. *Eur. J. Oper. Res.*, 223(2), 312–320. doi: 10.1016/j.ejor.2012.06.009.

Zlotnik, A., Chertkov, M., and Backhaus, S.N. (2015). Optimal control of transient flow in natural gas networks. *2015 54th IEEE Conference on Decision and Control (CDC)*, 4563–4570.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Evasive subspaces and rank-metric codes

Giuseppe Marino *

* *University of Naples "Federico II"(e-mail: giuseppe.marino@unina.it).*

Abstract: We investigate the connections between rank-metric codes and evasive $\mathbb{F}_q$-subspaces of $\mathbb{F}_{q^m}^k$. We show how the parameters of a rank-metric code are related to special geometric properties of the associated evasive subspace and construct new MRD-codes.

Keywords: Evasive subspace, MRD code, linear cutting blocking set, scattered subspace, q-polynomial

1. INTRODUCTION

Rank-metric codes, in particular MRD codes, have been studied since the 1970s and have seen much interest in recent years due to a wide range of applications including storage systems (Roth (1991)), cryptosystems (Gabidulin (1995)), spacetime codes (Lusina et al (2003)) and random linear network coding (Koetter et al (2008)).

In finite geometry, there are several interesting structures, including quasifields, semifields, splitting dimensional dual hyperovals and maximum scattered subspaces, which can be equivalently described as special types of rank-metric codes; see Csajbók et al (2017), Dempwolff et al (2014), Dempwolff et al (2015), Sheekey (2016), Taniguchi et al (2014) and the references therein.

2. RANK-METRIC CODES

The *rank weight* $\mathrm{rk}(v)$ of a vector $v = (v_1, \dots, v_n) \in \mathbb{F}_{q^m}^n$ is the dimension of the $\mathbb{F}_q$-linear space generated by its entries, i.e.

$$\mathrm{wt}_{\mathrm{rk}}(v) = \dim_{\mathbb{F}_q}(\langle v_1, \dots, v_n \rangle_{\mathbb{F}_q})$$

and the *rank distance* between two vectors is defined as $\mathrm{d}_{\mathrm{rk}}(u, v) := \mathrm{wt}_{\mathrm{rk}}(u - v)$.

Definition 1. An $[n, k, d]_{q^m/q}$ *(rank-metric) code* $\mathcal{C}$ is a k-dimensional $\mathbb{F}_{q^m}$-subspace of $\mathbb{F}_{q^m}^n$ equipped with the rank distance. The parameter d is called the *minimum rank distance* and it is given by

$$\begin{aligned} d := \mathrm{d}_{\mathrm{rk}}(\mathcal{C}) &= \min\{\mathrm{d}_{\mathrm{rk}}(u, v) \colon u, v \in \mathcal{C}, u \neq v\} \\ &= \min\{\mathrm{wt}_{\mathrm{rk}}(v) \colon v \in \mathcal{C}, v \neq 0\}. \end{aligned}$$

A *generator matrix* for $\mathcal{C}$ is a matrix $G \in \mathbb{F}_{q^m}^{k \times n}$ such that

$$\mathcal{C} = \{vG \colon v \in \mathbb{F}_{q^m}^k\}.$$

When the minimum distance is not known or is irrelevant, we write $[n, k]_{q^m/q}$.

It is well-known that

$$\#\mathcal{C} \leq q^{\max\{m,n\}(\min\{m,n\}-d+1)},$$

which is a Singleton like bound for the rank metric; see Delsarte (1978).

When equality holds, we call $\mathcal{C}$ a *maximum rank-distance* (MRD for short) code. More properties of MRD codes can be found in Delsarte (1978), Gabidulin (1985), Gadouleau (2006), Morrison (2014) and Ravagnani (2016).

There are several slightly different definitions of equivalence of rank-distance codes. Here, we use the following notion of equivalence: two $[n, k]_{q^m/q}$ codes $\mathcal{C}_1$ and $\mathcal{C}_2$ are *equivalent* if and only if there exist $A \in \mathrm{GL}(n, q)$ such that

$$\mathcal{C}_1 = \mathcal{C}_2 \cdot A = \{vA \colon v \in \mathcal{C}_2\}.$$

Definition 2. An $[n, k]_{q^m/q}$ code is *nondegenerate* if the $\mathbb{F}_q$-span of the columns of its generator matrices has $\mathbb{F}_q$-dimension equal to n.

Generalized rank weights have been first introduced and studied by Kurihara-Matsumoto-Uyematsu (Kurihara et al (2012),Kurihara et al (2015)), Oggier-Sboui Oggier et al (2012) and Ducoat Ducoat et al (2015).

Let $\Lambda_q(n, m)$ be the set of Frobeniuos-closed subspaces of $\mathbb{F}_{q^m}^n$ i.e.

$$\Lambda_q(n, m) := \{U \leq \mathbb{F}_{q^m}^n : \theta(U) = U\}\},$$

where $\theta : x \in \mathbb{F}_{q^m} \mapsto x^q \in \mathbb{F}_{q^m}$.

Definition 3. Let $\mathcal{C}$ be an $[n, k, d]_{q^m/q}$ code, and let $1 \leq s \leq k$. The *s-th generalized rank weight* of $\mathcal{C}$ is the integer

$$\mathrm{d}_{\mathrm{rk},s}(\mathcal{C}) := \min\{\dim_{\mathbb{F}_{q^m}}(\mathcal{A}) : \mathcal{A} \in \Lambda_q(n, m), \dim(\mathcal{A} \cap \mathcal{C}) \geq s\}.$$

Proposition 4. (Bounds Martínez-Peñas (2016)). Let $\mathcal{C}$ be an $[n, k, d]_{q^m/q}$ code and let $1 \leq s \leq k$. Then

$$\mathrm{d}_{\mathrm{rk},s}(\mathcal{C}) \leq \min\left\{n - k + s, sm, \frac{m}{n}(n - k) + m(s - 1) + 1\right\}.$$

Definition 5. Let s be a positive integer. An $[n, k]_{q^m/q}$ code $\mathcal{C}$ is s-MRD if $d_{\mathrm{rk},s}(\mathcal{C}) = n - k + s$.

According to Proposition 4, an $[n, k]_{q^m/q}$ s-MRD code may exist only if

$$n - k + s \leq \min\left\{sm, \frac{m}{n}(n - k) + m(s - 1) + 1\right\}.$$

3. EVASIVE SUBSPACES

Let $V = V(k, q^m)$ be a k-dimensional vector space over the finite field $\mathbb{F}_{q^m}$. Note that V is also a km-dimensional vector space over $\mathbb{F}_q$.

Definition 6. An $\mathbb{F}_q$-subspace U of V will be called $(h, r)_q$-*evasive* if $\langle U \rangle_{\mathbb{F}_{q^m}}$ has dimension at least h over $\mathbb{F}_{q^m}$ and the h-dimensional $\mathbb{F}_{q^m}$-subspaces of V meet U in $\mathbb{F}_q$-subspaces of dimension at most r. When $h = r$, an $(h, h)_q$-evasive subspace of V is said to be *h-scattered*.

Note that in the definition above the condition on the dimension of $\langle U \rangle_{\mathbb{F}_q^m}$ is to exclude trivial examples for which some of

our results would not apply. So, take an $\mathbb{F}_q$-subspace U of V such that the condition $\dim_{q^m}\langle U\rangle_{q^m} \geq h$ holds. Then it is easy to find an h-dimensional $\mathbb{F}_{q^m}$-subspace meeting U in an $\mathbb{F}_q$-subspace of dimension at least h and hence for an $(h,r)_q$-evasive subspace $h \leq r$ must hold. Clearly, if $\dim_q U \leq r$ then U is an $(h,r)_q$-evasive subspace. The $(k,r)_q$-evasive subspaces are the $\mathbb{F}_q$-subspaces of dimension at most r which span V over $\mathbb{F}_{q^m}$.

4. SOME BOUNDS OF EVASIVE SYSTEMS

Recently, the concept of q-systems has been introduced in Randrianarisoa (2020). A *q-system* U over $\mathbb{F}_{q^m}$ with parameters $[n,k,d]$ is an n-dimensional $\mathbb{F}_q$-subspace generating over $\mathbb{F}_{q^m}$ a k-dimensional $\mathbb{F}_{q^m}$-vector space V, where

$$d = n - \max\{\dim(U \cap H) : H \text{ is a hyperplane of } V\}.$$

With our notation, it is equivalent to say that U is $(k-1, n-d)_q$-evasive in $V(k,q^m)$ and it is not $(k-1, n-d+1)_q$-evasive.

Two $[n,k,d]_{q^m/q}$ systems U_1 and U_2 are said to be *(linearly) equivalent* if there exists $M \in \mathrm{GL}(k,q^m)$ such that $U_1 = M \cdot U_2$.

Let $\mathfrak{U}(n,k,d)_{q^m/q}$ denote the set of equivalence classes of $[n,k,d]_{q^m/q}$ systems, and let $\mathfrak{C}(n,k,d)_{q^m/q}$ denote the set of equivalence classes of nondegenerate $[n,k,d]_{q^m/q}$ codes. Then, define the maps

$$\begin{aligned} \Phi : \quad \mathfrak{C}(n,k,d)_{q^m/q} &\longrightarrow \mathfrak{U}(n,k,d)_{q^m/q} \\ [\mathrm{rowsp}(u_1^\top \mid \ldots \mid u_n^\top)] &\longmapsto [\langle u_1, \ldots u_n\rangle_{\mathbb{F}_q}] \end{aligned}$$

$$\begin{aligned} \Psi : \mathfrak{U}(n,k,d)_{q^m/q} &\longrightarrow \quad \mathfrak{C}(n,k,d)_{q^m/q} \\ [\langle u_1, \ldots u_n\rangle_{\mathbb{F}_q}] &\longmapsto [\mathrm{rowsp}(u_1^\top \mid \ldots \mid u_n^\top)] \end{aligned}$$

The maps Φ and Ψ are well-defined and they are inverse to each other (cf. Randrianarisoa (2020)).

The geometric point of view allows to give a geometric characterization of the generalized rank weights of a rank-metric code via its associated q-system.

Theorem 7. (Randrianarisoa (2020)). Let $\mathcal{C}$ be a nondegenerate $[n,k]_{q^m/q}$ code and let $U \in \Phi([\mathcal{C}])$ be any of the $[n,k]_{q^m/q}$ systems associated. Then

$$\begin{aligned} \mathrm{d}_{\mathrm{rk},r}(\mathcal{C}) = n - \max\{ &\dim_{\mathbb{F}_q}(U \cap W) : W \subseteq V, \\ &\dim_{\mathbb{F}_{q^m}}(W) = k - r\}. \end{aligned}$$

The following result gives a precise description of the connections between evasive subspaces and rank-metric codes.

Theorem 8. ((Marino et al, 2022, Thm. 3.3)). Let $\mathcal{C}$ be a nondegenerate $[n,k]_{q^m/q}$ code, and let $U \in \Phi([\mathcal{C}])$. Then, U is $(h,r)_q$-evasive if and only if $\mathrm{d}_{\mathrm{rk},k-h}(\mathcal{C}) \geq n-r$. In particular, $\mathrm{d}_{\mathrm{rk},k-h}(\mathcal{C}) = n-r$ if and only if U is $(h,r)_q$-evasive but not $(h,r-1)_q$-evasive.

Corollary 9. Let $\mathcal{C}$ be a nondegenerate $[n,k]_{q^m/q}$ code and let $U \in \Phi([\mathcal{C}])$. Then U is an h-scattered $[n,k]_{q^m/q}$ system if and only if $\mathcal{C}$ is $(k-h)$-MRD.

In Csajbók et al (2021), the following upper bound on the dimension of an h-scattered system has been proved.

Theorem 10. Let U be an h-scattered $[n,k]_{q^m/q}$ system. Then

$$n \leq \frac{km}{h+1}.$$

An $[n,k]_{q^m/q}$ system is said to be *maximum h-scattered* if it is h-scattered and meets the bound of Theorem 10 with equality, i.e.,

$$n = \frac{km}{h+1}.$$

Theorem 11. Let k, m be positive integers and let q be a prime power. The following hold.

(1) There exist maximum $(k-1)$-scattered $[m, k, m-k+1]_{q^m/q}$ systems (see Delsarte (1978)).
(2) If km is even, then there exist maximum 1-scattered $[\frac{km}{2}, k,]_{q^m/q}$ systems (see Blokhius et al (2000); Bartoli et al (2018); Csajbók et al (2017)).

In Zini et al (2021) a connection between maximum h-scattered subspaces and MRD codes has been established.

Theorem 12. ((Zini et al, 2021, Theorem 3.2)). Suppose that $h+1$ divides km and let $n := \frac{km}{h+1}$. Let U be an $[n,k]_{q^m/q}$ system and let $\mathcal{C} \in \Psi([U])$ be any of its associated $[n,k]_{q^m/q}$ codes. Then, U is maximum h-scattered if and only if $\mathcal{C}$ is an MRD code.

About the existence of maximum h-scattered the following result holds true.

Theorem 13. If $h+1$ divides k and $m \geq h+1$, then there exist maximum h-scattered $[\frac{km}{h+1}, k]_{q^m/q}$ systems.

If $h+1$ does not divide k we have the following result.

Theorem 14. If $m \geq 4$ is even and $r \geq 3$ is odd, then there exist maximum $(m-3)$-scattered $[\frac{rm}{2}, \frac{r(m-2)}{2}]_{q^m/q}$ system.

The main open problem about h-scattered in $V(k,q^m)$ is their existence for every admissible values of k, m and h. In particular the first open case is the existence of 2-scattered $\mathbb{F}_q$-subspaces of $V(4,q^6)$ of dimension 8.

From Bartoli et al (2021) and Marino et al (2022) we have the following more general bounds on evasive subspaces.

Proposition 15. Let U be a $(k-1,r)_q$-evasive $[n,k]_{q^m/q}$ system. Then

$$km \leq n(m-n+r+1).$$

Proposition 16. Let U be an $(h,r)_q$-evasive $[n,k]_{q^m/q}$ system. The following hold.

(1) $r \geq \max\{h, h-1+\frac{h+1}{m-1}\}$.

(2) $n \leq km - hm + r$,

(3) If $r \leq \frac{m}{m-1}h$, then

$$n \leq \frac{km}{r+1-m(r-h)}.$$

5. SMALL LINEAR CUTTING BLOCKING SETS AND NEW MRD

Another family of q-systems has been recently introduced in Alfarano et al. (2021).

Definition 17. An $[n,k]_{q^m/q}$ system U is said to be t-cutting if for every $\mathbb{F}_{q^m}$-subspace H of $V(k,q^m)$ of codimension t we have $\langle H \cap U\rangle_{\mathbb{F}_{q^m}} = H$. When $t=1$, we simply say that U is *cutting* (or a *linear cutting blocking set*).

The study of these objects was due to their connection to minimal rank-metric codes. A codeword is said to be *m*inimal if its support does not contain the support of any other non-proportional nonzero codeword. A code for which every codeword is minimal is called *minimal*. The importance of minimal codes relies on their connection with cryptography and coding theory.

Theorem 18. (Alfarano et al. (2021)). Let $\mathcal{C}$ be an $[n,k]_{q^m/q}$ code, and let $U \in \Phi([\mathcal{C}])$ be any of the associated $[n,k]_{q^m/q}$ systems. Then, $\mathcal{C}$ is a minimal rank-metric code if and only if U is a linear cutting blocking set.

The following bound on the parameters of linear cutting blocking set was determined in Alfarano et al. (2021).

Proposition 19. Let U be a cutting $[n,k]_{q^m/q}$ system, with $k \geq 2$. Then $n \geq m + k - 1$.

Moreover, in Alfarano et al. (2021) it was observed that linear cutting blocking sets are related with scattered subspaces when $k = 3$. In this case, scattered susbapces were used to construct linear cutting blocking sets, as the following result shows.

Proposition 20. (Alfarano et al. (2021)). If U is a scattered $[n,3]_{q^m/q}$ system with $n \geq m + 2$, then U is cutting.

The previous result has been generalized in Bartoli et al (in press).

Theorem 21. Let U be an $[n,k]_{q^m/q}$ system. Then, U is $(k-2, n-m-1)_q$-evasive if and only if it is cutting.

Note that, in general, the lower bound in Proposition 19 is not tight.

Corollary 22. If $m < (k-1)^2$ then there are no linear cutting blocking set of dimension $m + k - 1$.

Furthermore, it is interesting to observe what happens in the extremal case $m = (k-1)^2$.

Corollary 23. Let U be a $[k(k-1), k]_{q^{(k-1)^2}/q}$ system and let $\mathcal{C}$ be a code associated to U. The following are equivalent.

(1) U is $(k-2)$-scattered.
(2) U is cutting.
(3) $\mathcal{C}$ is MRD.
(4) $\mathcal{C}$ is minimal.

In Bartoli et al (in press) the cases $(k,m) = (4,3)$ and $(k,m) = (4,4)$ have been investigated.

Denote by $c_q(k,m)$ the smallest dimension of a linear cutting blocking set in $V(k,q^m)$. Then Proposition 19 can be rewritten as

$$c_q(k,m) \geq k + m - 1,$$

and we have seen that this is not always an equality; see Corollary 22. In Bartoli et al (in press) has been shown that $c_q(4,3) = 8$ for every prime power q. Also in Bartoli et al (in press), it has been proved that $c_q(4,4) \geq 8$. In particular it has shown that $c_q(4,4) = 8$ when $q = 2^{2h+1}$, with $h \geq 0$, constructing a scattered $[8,4]_{q^4/q}$ system which is also $(2,3)_q$-evasive. If $\mathcal{C}$ is an $[8,4]_{q^4/q}$ MRD associated with U, its generalized rank weights are

$$\mathrm{d}_{\mathrm{rk},1}(\mathcal{C}) = 3,\ \mathrm{d}_{\mathrm{rk},2}(\mathcal{C}) = 5,\ \mathrm{d}_{\mathrm{rk},3}(\mathcal{C}) = 7,\ \mathrm{d}_{\mathrm{rk},4}(\mathcal{C}) = 8.$$

The importance of this result is multiple, and it is related to the theory of rank-metric codes. On the one hand it provides a construction of a minimal rank-metric code of dimension 4 with respect to the field extension $\mathbb{F}_{q^4}/\mathbb{F}_q$ of shortest length. On the other hand it provides the first example of a $[2n, 2(n-d+1), d]_{q^n/q}$ MRD code which is not the direct sum of two $[n, n-d+1, d]_{q^n/q}$ MRD codes.

REFERENCES

G.N. Alfarano, M. Borello, A. Neri, and A. Ravagnani. Linear Cutting Blocking Sets and Minimal Codes in the Rank Metric. arXiv:2106.12465v1.

D. Bartoli, B. Csajbók, G. Marino, and R. Trombetti. Evasive subsapces. *J. Combin. Des. 2021*, 29, pages 533–551, 2021.

D. Bartoli, M. Giulietti, G. Marino, and O. Polverino. Maximum scattered linear sets and complete caps in Galois spaces. *Combinatorica*, 38 (2), pages 255–278, 2018.

D. Bartoli, G. Marino, and A. Neri. New MRD codes from linear cutting blocking sets. *Annali di Matematica Pura ed Applicata*, in press.

A. Blokhuis, and M. Lavrauw. Scattered spaces with respect to a spread in $\mathrm{PG}(n,q)$. *Geometriae Dedicata*, 81 (1), pages 231–243, 2000.

B. Csajbók, G. Marino, O. Polverino, and F. Zullo. Maximum scattered linear sets and MRD-codes. *J. Algebraic Combin.*, 46, pages 1–15, 2017.

B. Csajbók, G. Marino, O. Polverino, and F. Zullo. Generalising the scattered property of subspaces. *Combinatorica*, pages 1–26, 2021.

P. Delsarte. Bilinear forms over a finite field, with applications to coding theory. *Journal of Combinatorial Theory, Series A*, 25, pages 226–241, 1978.

U. Dempwolff, and Y. Edel. Dimensional dual hyperovals and APN functions with translation groups. *Journal of Algebraic Combinatorics*, 39, pages 457–496, 2014.

U. Dempwolff, and W. M. Kantor. Orthogonal dual hyperovals, symplectic spreads, and orthogonal spreads. *Journal of Algebraic Combinatorics*, 41, pages 83–108, 2015.

J. Ducoat, and G. Kyureghyan. *Generalized rank weights: a duality statement*. Topics in Finite Fields, 632, pages 101–109, 2015.

E.M. Gabidulin. Theory of codes with maximum rank distance. *Problems of information transmission*, 21, pages 3–16, 1985.

E.M. Gabidulin. Public-key cryptosystems based on linear codes over large alphabets: efficiency and weakness. In *Codes and Cyphers*, pages 17–31. Formara Limited, 1995.

M. Gadouleau, and Z. Yan. Properties of codes with the rank metric. In *IEEE Global Telecommunications Conference 2006*, pages 1–5, 2006.

R. Koetter, and F. Kschischang. Coding for errors and erasure in random network coding. *IEEE Transactions on Information Theory*, 54, pages 3579–3591, 2008.

J. Kurihara, T. Uyematsu, and R. Matsumoto. New parameters of linear codes expressing security performance of universal secure network coding. 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 533–540, 2012.

J. Kurihara, T. Uyematsu, and R. Matsumoto. Relative generalized rank weight of linear codes and its applications to network coding, *IEEE Transactions On information theory*, 61 (7), pages 3912–3936, 2015.

P. Lusina, E. Gabidulin, and M. Bossert. Maximum rank distance codes as space-time codes. *IEEE Transactions on Information Theory*, 49, pages 2757–2760, 2003.

G. Marino, A. Neri, and R. Trombetti. Evasive subspaces, generalized rank weights and near MRD codes. https://arxiv.org/abs/2204.11791.

U. Martínez-Peñas. On the similarities between generalized rank and Hamming weights and their applications to network coding. *IEEE Transactions on Information Theory*, 62, pages 4081–4095, 2016.

K. Morrison. Equivalence for rank-metric and matrix codes and automorphism groups of Gabidulin codes. *IEEE Transactions on Information Theory*, 60, pages 7035–7046, 2014.

F. Oggier, and A. Sboui. On the existence of generalized rank weights. International Symposium on Information Theory and its Applications, pages 406–410, 2012.

T. H. Randrianarisoa. A geometric approach to rank metric codes and a classification of constant weight codes. *Des. Codes Cryptogr.*, 88, pages 1331–1348, 2020.

A. Ravagnani. Rank-metric codes and their duality theory. *Designs, Codes and Cryptography*, 80, pages 197–216, 2016.

R.M. Roth. Maximum-rank array codes and their application to crisscross error correction. *IEEE Transactions on Information Theory*, 37, pages 328–336, 1991.

J. Sheekey. A new family of linear maximum rank distance codes. *Advances in Mathematics of Communications*, 10, pages 475–488, 2016.

H. Taniguchi, and S. Yoshiara. A unified description of four simply connected dimensional dual hyperovals. *European Journal of Combinatorics*, 36, pages 143–150, 2014.

G. Zini, and F. Zullo. Scattered subspaces and related codes. *Designs, Codes and Cryptography*, 89, pages 1853–1873, 2021.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Sampling formulae in reproducing kernel Hilbert spaces

Harry Dym *

* *Department of Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel (e-mail: harry.dym@weizmann.ac.il).*

Abstract: The classical sampling formula associated with the names of Shannon, Whittaker, Nyquist and Kotelnikov, will be exhibited as a special case of a general sampling formula in the setting of reproducing kernel Hilbert spaces of entire functions due to Louis de Branges. Other applications of these spaces and generalizations to spaces of vector valued entire functions will also be discussed briefly.

Keywords: Sampling formulas, reproducing kernel Hilbert spaces, de Branges spaces.

The classical sampling formula associated with the names of Shannon, Whittaker, Nyquist and Kotelnikov, may be formulated as follows:

If $f(\lambda) = \int_{-T}^{T} e^{i\mu s}\varphi(s)ds$ *for some function* φ *with the property* $\int_{-T}^{T} |\varphi(s)|^2 ds < \infty$, *then*

$$f(\lambda) = \sum_{n=-\infty}^{\infty} f(\mu_n)\frac{\sin(\lambda-\mu_n)T}{(\lambda-\mu_n)T} \quad \text{with } \mu_n = \frac{n\pi}{T}, \tag{1}$$

i.e., $f(\lambda)$ *can be recovered from its values at the points* μ_n.

This formula is usually derived by exploiting the interplay between the Fourier series expansion of φ and classical formulas for Fourier integrals.

I shall, however, discuss another approach that invokes the Paley-Wiener theorem to reinterpret this sampling formula as a special case of a family of such formulas that arise in a class of reproducing kernel Hilbert spaces of entire functions that was introduced and deeply investigated by Louis de Branges.

The de Branges theory begins with an entire function $E(\lambda)$ that is subject to the constraint

$$|E(\lambda)| > |E(\overline{\lambda})| \quad \text{for every point } \lambda \text{ in the open upper half-plane } \mathbb{C}_+. \tag{2}$$

To each such function $E(\lambda)$ there is an associated reproducing kernel Hilbert space $\mathcal{B}(E)$ of entire functions f (that may be characterized by some growth conditions on f) with reproducing kernel

$$K_\omega(\lambda) = \frac{E(\lambda)E(\omega)^* - E^{\#}(\lambda)E^{\#}(\omega)^*}{-2\pi i(\lambda-\overline{\omega})},$$

in which $h^{\#}(\lambda) = h(\overline{\lambda})^*$ for entire functions h, and inner product

$$\langle f, g\rangle_{\mathcal{B}(E)} = \int_{-\infty}^{\infty} (E^{-1}g)(\mu)^*(E^{-1}f)(\mu)d\mu.$$

This means that for each choice of $\omega \in \mathbb{C}$ and $f \in \mathcal{B}(E)$:

(i) $K_\omega \in \mathcal{B}(E)$.

(ii) $\langle f, K_\omega\rangle_{\mathcal{B}(E)} = f(\omega)$.

To proceed further, it is convenient to express the reproducing kernel in terms of the functions

$$A(\lambda) = \frac{E^{\#}(\lambda)+E(\lambda)}{2} = A^{\#}(\lambda)$$

and

$$B(\lambda) = \frac{E^{\#}(\lambda)-E(\lambda)}{2i} = B^{\#}(\lambda).$$

The constraint (2) forces the zeros of $B(\lambda)$, if any, to be real. Since

$$E(\lambda) = A(\lambda) - iB(\lambda) \quad \text{and} \quad E^{\#}(\lambda) = A(\lambda) + iB(\lambda),$$

it is readily checked that

$$K_\omega(\lambda) = \frac{A(\lambda)B(\overline{\omega}) - B(\lambda)A(\overline{\omega})}{\pi(\lambda-\overline{\omega})}$$

Moreover, if $\omega_1, \omega_2, \dots$ is a sequence of real zeros of $B(\lambda)$, then

$$\begin{aligned}\langle K_{\omega_j}, K_{\omega_k}\rangle_{\mathcal{B}(E)} &= K_{\omega_j}(\omega_k)\\ &= \frac{A(\omega_k)B(\omega_j) - B(\omega_k)A(\omega_j)}{\pi(\omega_k-\omega_j)}\\ &= 0 \quad \text{if } j \neq k.\end{aligned}$$

Under reasonable conditions on $E(\lambda)$, the set $\{K_{\omega_j}\}$ is an orthogonal basis for $\mathcal{B}(E)$ and hence every $f \in \mathcal{B}(E)$ can be expressed as

$$\begin{aligned}f(\lambda) &= \sum \langle f, K_{\omega_j}\rangle_{\mathcal{B}(E)} \|K_{\omega_j}\|^{-2}_{\mathcal{B}(E)} K_{\omega_j}(\lambda)\\ &= \sum f(\omega_j)\|K_{\omega_j}\|^{-2}_{\mathcal{B}(E)} K_{\omega_j}(\lambda).\end{aligned} \tag{3}$$

This is de Branges' sampling formula; see e.g., De Branges (1959), De Branges (1968); and, for an introduction to sampling theory Garcıa (2015).

Formula (1) is the special case of (3) that is obtained by choosing $E(\lambda) = e^{-i\lambda T}$, $T > 0$. For this choice of $E(\lambda)$, $A(\lambda) = \cos \lambda T$, $B(\lambda) = \sin \lambda T$ and

$$K_\omega(\lambda) = (\pi(\lambda-\overline{\omega}))^{-1}\sin(\lambda-\overline{\omega})T.$$

Formula (3) opens the door to many generalizations of (1).

The talk will be largely expository. However, if time permits, some highlights of a relatively recent paper, Dym

and Sarkar (2017) written jointly with S. Sarkar, that discusses sampling in de Branges spaces of vector valued entire functions, will be mentioned briefly; see also Arov and Dym (2018).

REFERENCES

Arov, D.Z. and Dym, H. (2018). Multivariate prediction, de branges spaces, and related extension and inverse problems. *Operator Theory: Advances and Applications*, volume 266, Birkhauser/Springer, Cham.

De Branges, L. (1959). Some mean squares of entire functions. *Proceedings of the American Mathematical Society*, 10(5), 833–839.

De Branges, L. (1968). *Hilbert spaces of entire functions*. Prentice-Hill, Englewood Cliffs, N.J.

Dym, H. and Sarkar, S. (2017). Multiplication operators with deficiency indices (p, p) and sampling formulas in reproducing kernel Hilbert spaces of entire vector valued functions. *Journal of Functional Analysis*, 273(12), 3671–3718.

Garcıa, A. (2015). Sampling theory and reproducing kernel Hilbert spaces. *Operator Theory*, 87–110.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Multilayer crisscross error and erasure correction

Umberto Martínez-Peñas *

* *IMUVa-Mathematics Research Institute, University of Valladolid, Spain (e-mail: umberto.martinez@uva.es).*

Abstract: In this work, the multi-cover metric is introduced. It is defined as a Cartesian product of classical cover metrics. A Singleton bound is given and maximum multi-cover distance (MMCD) codes are defined. Puncturing and shortening of linear MMCD codes are studied. It is shown that the dual of a linear MMCD code is not necessarily MMCD, and those satisfying this duality condition are defined as dually MMCD codes. Finally, constructions of dually MMCD codes are given, which also include some new linear codes attaining the Singleton bound for the classical cover metric and classical crisscross error correction.

Keywords: Crisscross error correction, duality, extremal codes, sum-rank distance

1. INTRODUCTION

The cover metric was introduced in (Roth, 1991) to measure the number of crisscross errors in memory chip arrays, which affect entire rows and columns. Two types of codes attaining the Singleton bound for the cover metric were introduced in (Roth, 1991), one of them based on the rank metric and patented in (Ordentlich et al., 2015).

In this work, we extend the cover metric to a tuple of ℓ matrices, where ℓ is a fixed positive integer. We call the new metric the multi-cover metric. This metric is suitable to correct simultaneously a number of crisscross errors and erasures distributed over the ℓ matrices.

In Section 2, we provide a Singleton bound for the multi-cover metric, and we call codes attaining it maximum multi-cover distance (MMCD) codes. As it can easily be seen, concatenated codes and Cartesian products of codes for the classical cover metric yield codes that are not MMCD. As in the sum-rank metric, MRD codes yield MMCD codes but only work for very tall matrices and are decodable over large finite fields. In Section 3, we study the puncturing and shortening of MMCD codes, and we show that the dual of a linear MMCD code is not necessarily MMCD, in contrast with MDS or MRD codes. Those satisfying this duality condition will be called dually MMCD codes. Finally, in Section 4, we provide a general family of dually MMCD codes for square matrices that can be corrected efficiently for a variety of alphabet sizes.

2. DEFINITIONS AND BASIC PROPERTIES

We denote $\mathbb{N} = \{0, 1, 2, \ldots\}$. Fix q a prime power and let $\mathbb{F}_q$ be the finite field of size q. For positive integers m and n, we denote by $\mathbb{F}_q^{m\times n}$ the set of $m \times n$ matrices with entries in $\mathbb{F}_q$. Throughout this manuscript, we will fix positive integers ℓ, $n_1, n_2, \ldots, n_\ell$, $m_1, m_2, \ldots, m_\ell$, and we will consider codes as subsets of $\prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$. We also set $\mathbf{m} = (m_1, m_2, \ldots, m_\ell)$ and $\mathbf{n} = (n_1, n_2, \ldots, n_\ell)$. We denote by $\mathrm{MC}(\mathbf{m}, \mathbf{n}) = \{(X_i, Y_i)_{i=1}^{\ell} \mid X_i \subseteq [m_i], Y_i \subseteq [n_i]\}$ the set of multi-covers in $\prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$. Finally, given $X = (X_i, Y_i)_{i=1}^{\ell} \in \mathrm{MC}(\mathbf{m}, \mathbf{n})$, we define its size as $|X| = \sum_{i=1}^{\ell}(|X_i| + |Y_i|)$ and its projection map $\pi_X : \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i} \longrightarrow \prod_{i=1}^{\ell} \mathbb{F}_q^{(m_i-|X_i|)\times(n_i-|Y_i|)}$ by removing from $C_i \in \mathbb{F}_q^{m_i\times n_i}$ the rows indexed by X_i and the columns indexed by Y_i in order to obtain $\pi_X(C_1, C_2, \ldots, C_\ell)$. Given $C \in \mathbb{F}_q^{m\times n}$ and $(C_1, C_2, \ldots, C_\ell) \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$, we denote by $C_{a,b}$ and $C_{i,a,b}$ the entry in row a and column b of the matrices C and C_i, respectively. Throughout the manuscript, we will also assume, without loss of generality, that $n_i \leq m_i$, for $i = 1, 2, \ldots, \ell$, and $m_1 \geq \ldots \geq m_\ell$.

We extend the definition of the cover metric from (Roth, 1991) to the multilayer case, as follows.

Definition 1. Let $C = (C_1, C_2, \ldots, C_\ell) \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$, where $C_i \in \mathbb{F}_q^{m_i\times n_i}$, for $i = 1, 2, \ldots, \ell$. We say that $(X_i, Y_i)_{i=1}^{\ell} \in \mathrm{MC}(\mathbf{m}, \mathbf{n})$ is a multi-cover of C if (X_i, Y_i) is a cover of C_i, which means that if $C_{i,a,b} \neq 0$, then $a \in X_i$ or $b \in Y_i$, for $i = 1, 2, \ldots, \ell$. We denote by $\mathrm{MC}(C)$ the set of all multi-covers of C. The multi-cover weight of C is then defined as $\mathrm{wt}_{MC}(C) = \min\{|X| \mid X \in \mathrm{MC}(C)\}$. The multi-cover metric is defined as $\mathrm{d}_{MC} : (\prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i})^2 \longrightarrow \mathbb{N}$, where $\mathrm{d}_{MC}(C, D) = \mathrm{wt}_{MC}(C - D)$, for $C, D \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$. Given a code $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$, we define its minimum multi-cover distance as $\mathrm{d}_{MC}(\mathcal{C}) = \min\{\mathrm{d}_{MC}(C, D) \mid C, D \in \mathcal{C}, C \neq D\}$. When considering the minimum distance of a code $\mathcal{C}$, we implicitly assume that $|\mathcal{C}| > 1$.

The cover metric (Roth, 1991) is recovered from the multi-cover metric when $\ell = 1$ (in particular, the multi-cover metric is indeed a metric as it is a sum of metrics). Similarly, both the multi-cover metric and the Hamming metric coincide in $\prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$ if $m_i = n_i = 1$, for $i = 1, 2, \ldots, \ell$. In this way, the multi-cover metric interpolates between the cover metric and the Hamming metric.

We conclude the section with a Singleton bound for the multi-cover metric. The proof is analogous to that of (Byrne et al., 2021, Th. 3.2) and is omitted.

Theorem 1. Let $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$ be a code, set $d = \mathrm{d}_{MC}(\mathcal{C})$ and let δ and j be the unique integers such that $d-1 = \sum_{i=1}^{j-1} n_i + \delta$ and $0 \leq \delta \leq n_j - 1$. Then

$$|\mathcal{C}| \leq q^{\sum_{i=j}^{\ell} m_i n_i - m_j \delta}. \tag{1}$$

In particular, if $m = m_1 = \ldots = m_\ell$ and if we set $N = n_1 + n_2 + \cdots + n_\ell$, then (1) reads

$$|\mathcal{C}| \leq q^{m(N-d+1)}.$$

We may thus define maximum multi-cover distance (MMCD) codes as follows.

Definition 2. We say that a code $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$ is maximum multi-cover distance (MMCD) if equality holds in (1).

3. DUALITY, PUNCTURING AND SHORTENING

In this section, we study duality, puncturing and shortening for the multi-cover metric. We will consider duality with respect to the trace product, given by

$$\langle C, D \rangle = \sum_{i=1}^{\ell} \mathrm{Tr}(C_i D_i), \tag{2}$$

where $C = (C_1, C_2, \ldots, C_\ell), D = (D_1, D_2, \ldots, D_\ell) \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$, and where $\mathrm{Tr}(A)$ denotes the trace of the matrix A. Observe that $\langle \cdot, \cdot \rangle$ is nothing but the usual inner product seeing $\prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$ as $\mathbb{F}_q^{\sum_{i=1}^{\ell} m_i n_i}$. We then define the dual of a linear code $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$ in the usual way,

$$\mathcal{C}^{\perp} = \left\{ D \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i} \;\middle|\; \langle C, D \rangle = 0, \text{ for all } C \in \mathcal{C} \right\}.$$

We now define puncturing and shortening, which extend the classical concepts of puncturing and shortening for the Hamming metric (Huffman and Pless, 2003, Sec. 1.5). As in the classical Hamming-metric case, puncturing and shortening enables us to explicitly construct shorter codes from known codes. To the best of our knowledge, these concepts have not been introduced in the classical cover metric case ($\ell = 1$).

Definition 3. Let $X = (X_i, Y_i)_{i=1}^{\ell} \in \mathrm{MC}(\mathbf{m}, \mathbf{n})$. Given a code $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$, we define its puncturing and shortening on X, respectively, as

$$\begin{aligned} \mathcal{C}_X &= \pi_X(\mathcal{C}), \\ \mathcal{C}^X &= \{\pi_X(C) \mid C \in \mathcal{C}, \text{ such that } C_{i,a,b} = 0 \\ &\qquad \text{if } a \in X_i \text{ or } b \in Y_i, 1 \leq i \leq \ell\}, \end{aligned}$$

both of which are codes in $\prod_{i=1}^{\ell} \mathbb{F}_q^{(m_i - |X_i|) \times (n_i - |Y_i|)}$.

We now describe the basic properties of punctured and shortened codes in general. The proof is left to the reader.

Proposition 4. Given a linear code $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$ of dimension k and minimum multi-cover distance d, and a multi-cover $X = (X_i, Y_i)_{i=1}^{\ell} \in \mathrm{MC}(\mathbf{m}, \mathbf{n})$, the following hold:

(1) $(\mathcal{C}_X)^{\perp} = (\mathcal{C}^{\perp})^X$ and $(\mathcal{C}^X)^{\perp} = (\mathcal{C}^{\perp})_X$.
(2) $\dim(\mathcal{C}_X) \geq k - \sum_{i=1}^{\ell} (n_i |X_i| + m_i |Y_i| - |X_i| \cdot |Y_i|)$.
(3) $\dim(\mathcal{C}^X) \geq k - \sum_{i=1}^{\ell} (n_i |X_i| + m_i |Y_i| - |X_i| \cdot |Y_i|)$.
(4) $\mathrm{d}_{MC}(\mathcal{C}_X) \geq d - |X|$.
(5) $\mathrm{d}_{MC}(\mathcal{C}^X) \geq d$.

More interestingly, we may obtain shorter linear MMCD codes from known linear MMCD codes.

Theorem 2. Let $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$ be a linear MMCD code. Set $d = \mathrm{d}_{MC}(\mathcal{C})$ and let δ and j be the unique integers such that $d-1 = \sum_{i=1}^{j-1} n_i + \delta$ and $0 \leq \delta \leq n_j - 1$. Let $X = (X_i, Y_i)_{i=1}^{\ell} \in \mathrm{MC}(\mathbf{m}, \mathbf{n})$. The following hold:

(1) Let $1 \leq k \leq j$, assume that $d > 1$, $X_i = \varnothing$, for all $i = 1, 2, \ldots, \ell$, $Y_i = \varnothing$ if $i \neq k$ and $|Y_k| = 1$. Further assume $\delta > 0$ if $k = j$. Then $\mathcal{C}_X$ is a linear MMCD code with $\dim(\mathcal{C}_X) = \dim(\mathcal{C})$ and $\mathrm{d}_{MC}(\mathcal{C}_X) = \mathrm{d}_{MC}(\mathcal{C}) - 1$.
(2) Let $j+1 \leq k \leq \ell$, and assume that $Y_i = \varnothing$, for all $i = 1, 2, \ldots, \ell$, $X_i = \varnothing$ if $i \neq k$ and $|X_k| = 1$. Then $\mathcal{C}^X$ is a linear MMCD code with $\dim(\mathcal{C}^X) = \dim(\mathcal{C}) - n_k$ and $\mathrm{d}_{MC}(\mathcal{C}^X) = \mathrm{d}_{MC}(\mathcal{C})$.
(3) Let $j+1 \leq k \leq \ell$, and assume that $X_i = \varnothing$, for all $i = 1, 2, \ldots, \ell$, $Y_i = \varnothing$ if $i \neq k$ and $|Y_k| = 1$. Then $\mathcal{C}^X$ is a linear MMCD code with $\dim(\mathcal{C}^X) = \dim(\mathcal{C}) - m_k$ and $\mathrm{d}_{MC}(\mathcal{C}^X) = \mathrm{d}_{MC}(\mathcal{C})$.

Proof. We start by proving Item 1. Let $n_i' = n_i$ if $i \neq k$, and let $n_k' = n_k - 1$. Note that $\mathcal{C}_X \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i'}$. Set also $\delta' = \delta - 1$ if $k = j$, and $\delta' = \delta$ otherwise. Thus $0 \leq \delta' \leq n_j' - 1$ by the assumptions. By Proposition 4, we have that

$$\mathrm{d}_{MC}(\mathcal{C}_X) - 1 \geq d - 2 = \sum_{i=1}^{j-1} n_i + \delta - 1 = \sum_{i=1}^{j-1} n_i' + \delta',$$

and since $d > 1$, we also have that

$$\dim(\mathcal{C}_X) = \dim(\mathcal{C}) = \sum_{i=j}^{\ell} m_i n_i - m_j \delta = \sum_{i=j}^{\ell} m_i n_i' - m_j \delta'.$$

Therefore, $\mathcal{C}_X$ must be MMCD and the inequalities above are equalities.

We next prove Item 2. Let $m_i' = m_i$ if $i \neq k$, and let $m_k' = m_k - 1$. Note that $\mathcal{C}^X \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i' \times n_i}$. By Proposition 4, we have that

$$\mathrm{d}_{MC}(\mathcal{C}^X) - 1 \geq d - 1 = \sum_{i=1}^{j-1} n_i + \delta, \text{ and}$$

$$\dim(\mathcal{C}^X) \geq \dim(\mathcal{C}) - n_k = \sum_{i=j}^{\ell} m_i' n_i - m_j' \delta.$$

Therefore, $\mathcal{C}^X$ must be MMCD and the inequalities above are equalities.

Finally we prove Item 3. Let $n_i' = n_i$ if $i \neq k$, and let $n_k' = n_k - 1$. Note that $\mathcal{C}^X \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i'}$. By Proposition 4, we have that

$$\mathrm{d}_{MC}(\mathcal{C}^X) - 1 \geq d - 1 = \sum_{i=1}^{j-1} n_i + \delta = \sum_{i=1}^{j-1} n_i' + \delta, \text{ and}$$

$$\dim(\mathcal{C}^X) \geq \dim(\mathcal{C}) - m_k = \sum_{i=j}^{\ell} m_i n_i' - m_j \delta.$$

Therefore, $\mathcal{C}^X$ must be MMCD and the inequalities above are equalities. □

We now discuss duality and show that, in general, the dual of a linear MMCD code is not necessarily MMCD, leading to the concept of dually MMCD codes. We also relate the duality of MMCD codes with codes which are MDS by rows and columns.

Define the column-wise Hamming weight and metric in $\mathbb{F}_q^{m\times n}$ in the natural way and denote them, respectively, by wt_H^C and d_H^C. In other words, $\mathrm{wt}_H^C(C)$ is the number of non-zero columns in the matrix $C \in \mathbb{F}_q^{m\times n}$. We may similarly define the Hamming metric by rows in $\mathbb{F}_q^{m\times n}$ by transposition, i.e., we may consider

$$\mathcal{C}^\intercal = \{C^\intercal \mid C \in \mathcal{C}\} \subseteq \mathbb{F}_q^{m\times n},$$

for $\mathcal{C} \subseteq \mathbb{F}_q^{n\times m}$, where $C^\intercal$ denotes the transposed of a matrix C. That is, $\mathrm{wt}_H^R(C) = \mathrm{wt}_H^C(C^\intercal)$. We denote the row-wise Hamming weight and metric in $\mathbb{F}_q^{m\times n}$ by wt_H^R and d_H^R, respectively. We denote the minimum Hamming distance of $\mathcal{C}$ by rows and by columns, respectively, by $\mathrm{d}_H^R(\mathcal{C})$ and $\mathrm{d}_H^C(\mathcal{C})$. Clearly, for $C \in \mathbb{F}_q^{m\times n}$ and $\mathcal{C} \subseteq \mathbb{F}_q^{m\times n}$,

$$\mathrm{wt}_{MC}(C) \leq \mathrm{wt}_H^C(C) \quad \text{and} \quad \mathrm{d}_{MC}(\mathcal{C}) \leq \mathrm{d}_H^C(\mathcal{C}), \tag{3}$$

and analogously for the row-wise Hamming weight and metric.

Consider now the square case $m = n$. The Singleton bound from Theorem 1 holds in the same way for both rows and columns, i.e.,

$$|\mathcal{C}| \leq q^{n(n-d+1)},$$

whether $d = \mathrm{d}_H^R(\mathcal{C})$ or $d = \mathrm{d}_H^C(\mathcal{C})$. A code attaining this bound for d_H^R will be called MDS by rows, and analogously for columns. Clearly, if $\mathcal{C}$ is MMCD, then it is both MDS by rows and by columns. In particular, if it is linear, then $\mathcal{C}^\perp$ is also MDS by rows and by columns, since the MDS property is preserved by duality (Huffman and Pless, 2003, Th. 2.4.3).

However, as we now show, the dual of a linear MMCD code is not necessarily MMCD itself and a linear code that is MDS by rows and by columns is not necessarily MMCD either.

Example 5. Consider $\mathcal{C} \subseteq \mathbb{F}_2^{3\times 3}$ generated by

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \text{and} \quad C = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Since $\mathcal{C} = \{0, A, B, C, A+B, B+C, C+A, A+B+C\}$ and $\dim(\mathcal{C}) = 3$, it is easy to see that $\mathcal{C}$ is MDS by columns and by rows since $\mathrm{d}_H^R(\mathcal{C}) = \mathrm{d}_H^C(\mathcal{C}) = 3$, but it is not MMCD, since $\mathrm{d}_{MC}(\mathcal{C}) = 2$. Now, $\mathcal{C}^\perp \subseteq \mathbb{F}_2^{3\times 3}$ has $\dim(\mathcal{C}^\perp) = 6$. Moreover, by inspection one can see that there is no $D \in \mathcal{C}^\perp$ with $\mathrm{wt}_{MC}(D) = 1$. Therefore, $\mathrm{d}_{MC}(\mathcal{C}^\perp) = 2$ and $\mathcal{C}^\perp$ is a linear MMCD code, even though $\mathcal{C}$ is not.

The example above motivates the following definition.

Definition 6. We say that $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$ is a dually MMCD code if it is linear and both $\mathcal{C}$ and $\mathcal{C}^\perp$ are MMCD codes.

In Section 4, we will provide some explicit constructions of dually MMCD codes for general parameters.

To conclude the subsection, we observe that the equivalence of linear MMCD codes, dually MMCD codes and MDS codes by rows and by columns holds for very small parameters. For the case $\ell > 1$, we may extend column-wise and row-wise Hamming weights and metrics to $\prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$ in a straightforward way. However, we also need to consider different combinations of transpositions in different positions. To this end, given $C = (C_1, C_2, \ldots, C_\ell) \in \prod_{i=1}^{\ell} \mathbb{F}_q^{n_i\times n_i}$ and $\mathbf{t} \in \{0,1\}^\ell$, we define $C^{\mathbf{t}} = (D_1, D_2, \ldots, D_\ell) \in \prod_{i=1}^{\ell} \mathbb{F}_q^{n_i\times n_i}$, where

$$D_i = \begin{cases} C_i & \text{if } t_i = 0, \\ C_i^\intercal & \text{if } t_i = 1. \end{cases}$$

We then define $\mathcal{C}^{\mathbf{t}} = \{C^{\mathbf{t}} \mid C \in \mathcal{C}\} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{n_i\times n_i}$, for $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{n_i\times n_i}$. Notice that the bounds (3) still hold for $\ell > 1$ and any vector of transpositions $\mathbf{t} \in \{0,1\}^\ell$.

The proofs of the following two propositions are based on the fact that, in the two cases, we only need to consider covers in $\mathbb{F}_q^{m\times n}$ of sizes 1 or 2, and these can always be chosen as only columns or only rows in such cases.

Proposition 7. Let $\mathcal{C} \subseteq (\mathbb{F}_q^{n\times n})^\ell$ be a linear code with $\dim(\mathcal{C}) = n(\ell n - 1)$. Then $\mathcal{C}$ is MMCD if, and only if, $\mathcal{C}^{\mathbf{t}}$ is MDS by columns for all $\mathbf{t} \in \{0,1\}^\ell$.

Proposition 8. Let $\mathcal{C} \subseteq (\mathbb{F}_q^{2\times 2})^\ell$ be a linear code. The following are equivalent:

(1) $\mathcal{C}^{\mathbf{t}}$ is MDS by columns for all $\mathbf{t} \in \{0,1\}^\ell$.
(2) $\mathcal{C}$ is MMCD.
(3) $\mathcal{C}$ is dually MMCD.

Recall that the multi-cover metric in $(\mathbb{F}_q^{1\times 1})^\ell$ is simply the classical Hamming metric in $\mathbb{F}_q^\ell$, hence the previous proposition also holds but is trivial in this case.

4. DUALLY MMCD CODE CONSTRUCTIONS

4.1 MSRD codes

Similarly to the case of the rank metric and the cover metric (Roth, 1991), we show in this subsection that sum-rank metric codes may be used as multi-cover metric codes. The sum-rank metric was formally defined in (Nóbrega and Uchôa-Filho, 2010, Sec. III-D), but it was implicitly used earlier in (Lu and Kumar, 2005, Sec. III).

Definition 9. We define the sum-rank weight of $C = (C_1, C_2, \ldots, C_\ell) \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$ as

$$\mathrm{wt}_{SR}(C) = \sum_{i=1}^{\ell} \mathrm{Rk}(C_i).$$

The sum-rank metric is then defined by $\mathrm{d}_{SR}(C, D) = \mathrm{wt}_{SR}(C - D)$, for $C, D \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i\times n_i}$.

The bound in Theorem 1 is also valid for the sum-rank metric (Byrne et al., 2021, Th. III.2), and a code attaining it is called maximum sum-rank distance (MSRD).

We have the following connections between both metrics. They constitute a trivial extension to $\ell \geq 1$ of the corresponding results for the case $\ell = 1$, observed in (Roth, 1991). We note that Item 4 follows from combining Item 3 with the fact that the dual of a linear MSRD code is again

$$\varphi\left(C^1, C^2, \ldots, C^r\right) = \left(\begin{array}{cccc|cccc|c|cccc} C_1^1 & C_1^2 & \ldots & C_1^r & C_{r+1}^1 & C_{r+1}^2 & \ldots & C_{r+1}^r & \ldots & C_{(\ell-1)r+1}^1 & C_{(\ell-1)r+1}^2 & \cdots & C_{(\ell-1)r+1}^r \\ C_2^r & C_2^1 & \ldots & C_2^{r-1} & C_{r+2}^r & C_{r+2}^1 & \ldots & C_{r+2}^{r-1} & \ldots & C_{(\ell-1)r+2}^r & C_{(\ell-1)r+2}^1 & \cdots & C_{(\ell-1)r+2}^{r-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ C_r^2 & C_r^3 & \ldots & C_r^1 & C_{2r}^2 & C_{2r}^3 & \ldots & C_{2r}^1 & \ldots & C_t^2 & C_t^3 & \ldots & C_t^1 \end{array}\right).$$

MSRD code under the given conditions (Byrne et al., 2021, Th. VI.1).

Proposition 10. Fix $C \in \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$ and $\mathcal{C} \subseteq \prod_{i=1}^{\ell} \mathbb{F}_q^{m_i \times n_i}$. The following hold:

(1) $\mathrm{wt}_{SR}(C) \leq \mathrm{wt}_{MC}(C)$.
(2) $\mathrm{d}_{SR}(\mathcal{C}) \leq \mathrm{d}_{MC}(\mathcal{C})$.
(3) If $\mathcal{C}$ is an MSRD code, then it is also MMCD.
(4) If $\mathcal{C}$ is a linear MSRD code and $m_1 = m_2 = \ldots = m_\ell$, then $\mathcal{C}$ is a dually MMCD code.

4.2 A nested construction

In this subsection, we provide a general method to construct codes for the multi-cover metric from other multi-cover metric codes. For simplicity, we will only work with square matrices and tuples of matrices with the same number of rows and columns, i.e., $m_1 = \ldots = m_\ell = n_1 = \ldots = n_\ell$. Throughout this subsection, we fix positive integers $n = rs$ and $t = r\ell$.

Construction 1. Let $\mathcal{C} \subseteq (\mathbb{F}_q^{s\times s})^t$ be a code. We define another code $\varphi(\mathcal{C}) \subseteq (\mathbb{F}_q^{n\times n})^\ell$ as the image of the linear map $\varphi : ((\mathbb{F}_q^{s\times s})^t)^r \longrightarrow (\mathbb{F}_q^{n\times n})^\ell$, where $\varphi\left(C^1, C^2, \ldots, C^r\right)$ is described at the top of this page, for $C^i = (C_1^i, C_2^i, \ldots, C_t^i) \in (\mathbb{F}_q^{s\times s})^t$, for $i = 1, 2, \ldots, r$.

We now relate the multi-cover metric parameters of $\mathcal{C}$ and $\varphi(\mathcal{C})$. The proof is left to the reader.

Theorem 3. Let $\mathcal{C} \subseteq (\mathbb{F}_q^{s\times s})^t$. The following hold:

(1) $\mathrm{d}_{MC}(\varphi(\mathcal{C})) = \mathrm{d}_{MC}(\mathcal{C})$ and $|\varphi(\mathcal{C})| = |\mathcal{C}|^r$.
(2) $\varphi(\mathcal{C})$ is MMCD if, and only if, so is $\mathcal{C}$.
(3) $\varphi(\mathcal{C})$ is linear if, and only if, so is $\mathcal{C}$, and in that case, $\dim(\varphi(\mathcal{C})) = r\dim(\mathcal{C})$ and $\varphi(\mathcal{C})^\perp = \varphi(\mathcal{C}^\perp)$.
(4) (If $\mathcal{C}$ is linear) $\varphi(\mathcal{C})$ is a dually MMCD code if, and only if, so is $\mathcal{C}$.

In this way, we may construct codes in $(\mathbb{F}_q^{n\times n})^\ell$ for the multi-cover metric from codes in $(\mathbb{F}_q^{s\times s})^t$ for the refined multi-cover metric. Note that, if we set $s = 1$, then we may construct multi-cover metric codes in $(\mathbb{F}_q^{n\times n})^\ell$ from Hamming-metric codes in $\mathbb{F}_q^t$.

4.3 Explicit linear MMCD codes

In this subsection, we put together the two methods for constructing multi-cover metric codes from Subsections 4.1 and 4.2 in order to give an explicit family of MMCD codes. We consider linearized Reed-Solomon codes (Martínez-Peñas, 2018) as component codes in Construction 1, for several choices of the integer t.

Theorem 4. Let $n = rs$ and $t = r\ell$ be positive integers, such that $q > t$. Let $\mathcal{C} \subseteq (\mathbb{F}_q^{s\times s})^t$ be a linearized Reed-Solomon code (Martínez-Peñas, 2018, Def. 31). Then the code $\varphi(\mathcal{C}) \subseteq (\mathbb{F}_q^{n\times n})^\ell$ obtained from $\mathcal{C}$ as in Construction 1, is an MMCD code. Using an r-folded version of the decoder from (Martínez-Peñas and Kschischang, 2019), it may correct errors of multi-cover weight at most $\lfloor(\mathrm{d}_{MC}(\mathcal{C}) - 1)/2\rfloor$ with a complexity of $\mathcal{O}(t\ell n^2)$ sums and products over the finite field of size $q^{\ell n/t} = \mathcal{O}(t)^{\ell n/t}$.

Assume that a multiplication in $\mathbb{F}_{2^b}$ costs $\mathcal{O}(b^2)$ operations in $\mathbb{F}_2$. Then if q is even, the MMCD code in Theorem 4 can be decoded with a complexity of

$$\mathcal{O}\left(t^{-1}\log_2(t+1)^2\ell^3 n^4\right)$$

operations over $\mathbb{F}_2$. This complexity is smaller for larger values of t. However, the alphabet size for the multi-cover metric needs to satisfy $q > t$. Thus we arrive at an alphabet-complexity trade-off: Codes for larger t are faster to decode but require larger alphabets, whereas codes for smaller t are less fast but can be used for a wider range of alphabets.

Finally, observe that, if $\ell = 1$ (thus $1 \leq t = r \leq n$), then the codes in Theorem 4 attain the Singleton bound for the classical cover metric (Roth, 1991). The cases $t = r = 1$ and $t = r = n$ were obtained already in (Roth, 1991), but the cases $1 < t = r < n$ of such codes constitute a new family of codes attaining the Singleton bound for the classical cover metric. Their interest resides in the alphabet-complexity trade-off mentioned above.

REFERENCES

Byrne, E., Gluesing-Luerssen, H., and Ravagnani, A. (2021). Fundamental properties of sum-rank-metric codes. *IEEE Trans. Info. Theory*, 67(10), 6456–6475.

Huffman, W.C. and Pless, V. (2003). *Fundamentals of error-correcting codes.* Cambridge University Press, Cambridge.

Lu, H.F. and Kumar, P.V. (2005). A unified construction of space–time codes with optimal rate-diversity tradeoff. *IEEE Trans. Info. Theory*, 51(5), 1709–1730.

Martínez-Peñas, U. (2018). Skew and linearized Reed–Solomon codes and maximum sum rank distance codes over any division ring. *J. Algebra*, 504, 587–612.

Martínez-Peñas, U. and Kschischang, F.R. (2019). Reliable and secure multishot network coding using linearized Reed-Solomon codes. *IEEE Trans. Info. Theory*, 65(8), 4785–4803.

Nóbrega, R.W. and Uchôa-Filho, B.F. (2010). Multishot codes for network coding using rank-metric codes. In *Proc. Third IEEE Int. Workshop Wireless Network Coding*, 1–6.

Ordentlich, E., Roth, R.M., and Seroussi, G. (2015). Memory controller using crisscross error-correcting codes. US Patent 9,070,436.

Roth, R.M. (1991). Maximum-rank array codes and their application to crisscross error correction. *IEEE Trans. Info. Theory*, 37(2), 328–336.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Reduced order models from data via generalized balanced truncation ⋆

Azka M. Burohman [*,**,***] **B. Besselink** [*,***]
Jacquelien M.A. Scherpen [**,***] **M. Kanat Camlibel** [*,***]

[*] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, The Netherlands,*
[**] *Engineering and Technology Institute Groningen (ENTEG), University of Groningen, Groningen, The Netherlands.*
[***] *Jan C. Willems Center for Systems and Control, University of Groningen, The Netherlands (e-mail: a.m.burohman@rug.nl, b.besselink@rug.nl, j.m.a.scherpen@rug.nl, m.k.camlibel@rug.nl).*

Keywords: Data-driven model reduction, generalized balancing, error bounds

1. INTRODUCTION

Since recently, the problem of *data-driven* model reduction is attracting increasing attention, partly motivated by the widespread availability of measurement data. Here, low-order models are constructed directly on the basis of measurement data, thus not requiring the availability of a high-order model. Nevertheless, standard (model-based) model reduction techniques have inspired various data-driven techniques.

In the class of energy-based methods for linear systems, to which this paper belongs, data-driven model reduction contributions include Rapisarda and Trentelman (2011); Markovsky et al. (2005); Gosea et al. (2022). Meanwhile, in the class of interpolatory methods, the contributions include Beattie and Gugercin (2012); Peherstorfer et al. (2017); Scarciotti and Astolfi (2017); Burohman et al. (2020).

Despite these developments, existing methods for data-driven model reduction do often not allow for guaranteeing system properties such as asymptotic stability and do not provide an error bound, especially when the available data is subject to noise. In this extended abstract, we develop a data-driven reduction technique that provides such guarantees on the low-order model, even for noisy data.

This work has the following contributions. First, we introduce the concept of *data reduction* via a Petrov-Galerkin projection. Following the data informativity framework of van Waarde et al. (2020), a class of systems explaining the data can be characterized by a quadratic matrix matrix inequality (QMI). Then, we show that projection of all systems in this class results in a class of reduced-order models that can also be characterized by a QMI, but one of smaller dimension.

The second contribution of this paper is the development of a data-driven generalized balanced truncation method. Here, we first give necessary and sufficient conditions for all systems explaining the data to have a *common* generalized controllability and *common* generalized observability Gramians. In this case, we say that the data are *informative* for generalized Lyapunov balancing. Next, we comprise the use of the common generalized Gramians to obtain the Petrov-Galerkin projection that achieves (generalized) balanced truncation (see Dullerud and Paganini (2000). This allows for the application of the data reduction concept.

As inherent advantages of using a balancing-type reduction method, all reduced-order models are guaranteed to be asymptotically stable and satisfy an *a priori* error bound. However, the ordinary a priori upper bounds from model-based reduction methods, e.g., Antoulas (2005); Dullerud and Paganini (2000), do not determine the error between a selected reduced-order model to the true system generating the data as the true system is unknown. Therefore, as the final contribution, we compute a uniform a priori upper bound, i.e., an error bound that holds for any chosen high-order system explaining the data and any reduced-order model.

The proofs of the results presented in this abstract can be found in Burohman et al. (2021).

⋆ This extended abstract is based on research developed in the DSSC Doctoral Training Programme, co-funded through a Marie Sklodowska-Curie COFUND (DSSC 754315).

2. PRELIMINARIES

2.1 Model reduction via a Petrov-Galerkin projection

Consider the discrete-time input/state/output system

$$\Sigma : \quad \begin{aligned} \boldsymbol{x}(k+1) &= A\boldsymbol{x}(k) + B\boldsymbol{u}(k), \\ \boldsymbol{y}(k) &= C\boldsymbol{x}(k) + D\boldsymbol{u}(k), \end{aligned} \tag{1}$$

with input $\boldsymbol{u} \in \mathbb{R}^m$, state $\boldsymbol{x} \in \mathbb{R}^n$ and output $\boldsymbol{y} \in \mathbb{R}^p$. Let $\hat{W}, \hat{V} \in \mathbb{R}^{n\times r}$ be matrices such that $\hat{W}^\top \hat{V} = I$ and $r < n$. A reduced-order model (ROM) obtained via a Petrov-Galerkin projection is given by

$$\hat{\Sigma} : \quad \begin{aligned} \hat{\boldsymbol{x}}(k+1) &= \hat{W}^\top A\hat{V}\hat{\boldsymbol{x}}(k) + \hat{W}^\top B\boldsymbol{u}(k), \\ \hat{\boldsymbol{y}}(k) &= C\hat{V}\hat{\boldsymbol{x}}(k) + D\boldsymbol{u}(k) \end{aligned} \tag{2}$$

where $\hat{\boldsymbol{x}} \in \mathbb{R}^r$ and $\hat{\boldsymbol{y}} \in \mathbb{R}^p$ denote the state and output of the ROM, respectively.

2.2 Generalized Lyapunov balancing

Consider the discrete-time system (1). Then, matrices $P = P^\top > 0$ satisfying

$$APA^\top - P + BB^\top < 0$$

and $Q = Q^\top > 0$ satisfying

$$A^\top QA - Q + C^\top C < 0,$$

are called the *generalized controllability Gramian* and *generalized observability Gramian*, respectively. This is a strict version of the definition of generalized Gramians given in Dullerud and Paganini (2000). The generalized Gramians are lower bounded by the ordinary Gramians, i.e., $P > P_0$ and $Q > Q_0$, where Q_0 and P_0 are the solutions of the corresponding Lyapunov *equations*.

As ordinary Gramians, these generalized Gramians can be used to obtain a so-called balanced realization. Specifically, by e.g., (Antoulas, 2005, Lemma 7.3), there exists a nonsingular matrix T such that $TPT^\top = T^{-\top}QT^{-1} = \Sigma_H$ where Σ_H is a diagonal matrix of the *generalized* Hankel singular values (GHSVs) of Σ in (1), i.e.,

$$\Sigma_H := \mathrm{blkdiag}(\sigma_1 I_{m_1}, \sigma_2 I_{m_2}, \ldots, \sigma_\kappa I_{m_\kappa}), \tag{3}$$

where $\sigma_1 > \sigma_2 > \cdots > \sigma_\kappa > 0$, and m_i denotes the multiplicity of σ_i for $i = 1, \ldots, \kappa$ satisfying $n = \sum_{i=1}^{\kappa} m_i$. In particular, the balanced realization is given by $A_{\mathrm{bal}} := TAT^{-1}$, $B_{\mathrm{bal}} := TB$, $C_{\mathrm{bal}} := CT^{-1}$, $D_{\mathrm{bal}} := D$. Moreover, similar to the standard balanced truncation, the corresponding model reduction error is twice the sum of the neglected GHSVs.

3. DATA-DRIVEN PETROV-GALERKIN PROJECTION

Consider the linear discrete-time input/state/output system

$$\Sigma_{\mathrm{true}}: \quad \begin{aligned} \boldsymbol{x}(k+1) &= A_{\mathrm{true}}\boldsymbol{x}(k) + B_{\mathrm{true}}\boldsymbol{u}(k) + \boldsymbol{w}(k), \\ \boldsymbol{y}(k) &= C_{\mathrm{true}}\boldsymbol{x}(k) + D_{\mathrm{true}}\boldsymbol{u}(k) + \boldsymbol{z}(k), \end{aligned} \tag{4}$$

where $(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{m+n+p}$ are the input/state/output and $(\boldsymbol{w}, \boldsymbol{z}) \in \mathbb{R}^{n+p}$ are noise terms. Throughout the paper, we assume that the system matrices $(A_{\mathrm{true}}, B_{\mathrm{true}}, C_{\mathrm{true}}, D_{\mathrm{true}})$ and the noise $(\boldsymbol{w}, \boldsymbol{z})$ are *unknown*. What is known instead are a finite number of input/state/output measurements harvested from the true system (4) collected in the matrices

$$\begin{aligned} X_- &:= [x(0)\ x(1)\ \cdots\ x(L-1)], \\ X_+ &:= [x(1)\ x(2)\ \cdots\ x(L)], \\ U_- &:= [u(0)\ u(1)\ \cdots\ u(L-1)], \\ Y_- &:= [y(0)\ y(1)\ \cdots\ y(L-1)]. \end{aligned}$$

Now, we can define the set of all systems that *explain* the data as

$$\boldsymbol{\Sigma} := \left\{ (A, B, C, D) : \begin{bmatrix} X_+ \\ Y_- \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X_- \\ U_- \end{bmatrix} \in \mathcal{N} \right\},$$

where $\mathcal{N} \subseteq \mathbb{R}^{(n+p)\times L}$ captures a *noise model* such that

$$(A_{\mathrm{true}}, B_{\mathrm{true}}, C_{\mathrm{true}}, D_{\mathrm{true}}) \in \boldsymbol{\Sigma}. \tag{5}$$

In this paper, we work with a noise model that is described by a quadratic matrix inequality as

$$\mathcal{N} := \left\{ Z \in \mathbb{R}^{(n+p)\times L} : \begin{bmatrix} I \\ Z^\top \end{bmatrix}^\top \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{12}^\top & \Phi_{22} \end{bmatrix} \begin{bmatrix} I \\ Z^\top \end{bmatrix} \succcurlyeq 0 \right\}, \tag{6}$$

where $\Phi_{11} = \Phi_{11}^\top \in \mathbb{R}^{(n+p)\times(n+p)}$, $\Phi_{12} \in \mathbb{R}^{(n+p)\times L}$, and $\Phi_{22} = \Phi_{22}^\top \in \mathbb{R}^{L\times L}$.

We make the following blanket assumption on the set $\mathcal{N}$.

Assumption 1. The set $\mathcal{N}$ is bounded and has nonempty interior.

As shown in van Waarde et al. (2022b), Assumption 1 holds if and only if $\Phi_{11} - \Phi_{12}\Phi_{22}^{-1}\Phi_{12}^\top > 0$ and $\Phi_{22} < 0$. It is clear from the definition of $\boldsymbol{\Sigma}$ and (6) that $(A, B, C, D) \in \boldsymbol{\Sigma}$ if and only if the following quadratic matrix inequality (QMI) is satisfied

$$\begin{bmatrix} I & 0 \\ 0 & I \\ A^\top & C^\top \\ B^\top & D^\top \end{bmatrix}^\top N \begin{bmatrix} I & 0 \\ 0 & I \\ A^\top & C^\top \\ B^\top & D^\top \end{bmatrix} \succcurlyeq 0, \tag{7}$$

where

$$N := \begin{bmatrix} I & 0 & X_+ \\ 0 & I & Y_- \\ 0 & 0 & -X_- \\ 0 & 0 & -U_- \end{bmatrix} \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{12}^\top & \Phi_{22} \end{bmatrix} \begin{bmatrix} I & 0 & X_+ \\ 0 & I & Y_- \\ 0 & 0 & -X_- \\ 0 & 0 & -U_- \end{bmatrix}^\top. \tag{8}$$

As a first step, we consider a Petrov-Galerkin projection and assuming that the projection matrices $\hat{W}$ and $\hat{V}$ satisfying $\hat{W}^\top\hat{V} = I$ are given. Then, the set of reduced-order models of systems explaining the data is defined as

$$\boldsymbol{\Sigma}_{\hat{V},\hat{W}} := \left\{ (\hat{W}^\top A\hat{V}, \hat{W}^\top B, C\hat{V}, D) : (A, B, C, D) \in \boldsymbol{\Sigma} \right\}.$$

The first main result of this paper is that the set $\boldsymbol{\Sigma}_{\hat{V},\hat{W}}$ can itself be represented as a QMI of a similar form as (7).

Theorem 1. Consider the set $\boldsymbol{\Sigma}$ of systems explaining the data. Suppose that there exists $\bar{S}$ such that

$$\begin{bmatrix} I \\ \bar{S} \end{bmatrix}^\top N \begin{bmatrix} I \\ \bar{S} \end{bmatrix} > 0 \tag{9}$$

holds and the matrix $\begin{bmatrix} X_- \\ U_- \end{bmatrix}$ has full row rank. Let $\hat{W}, \hat{V} \in \mathbb{R}^{n\times r}$ be such that $\hat{W}^\top\hat{V} = I$. Then, the set $\boldsymbol{\Sigma}_{\hat{V},\hat{W}}$ of reduced-order models of $\boldsymbol{\Sigma}$ using projection matrices $\hat{W}, \hat{V}$ satisfies

$$\boldsymbol{\Sigma}_{\hat{V},\hat{W}} = \left\{ (\hat{A}, \hat{B}, \hat{C}, \hat{D}) : \begin{bmatrix} I & 0 \\ 0 & I \\ \hat{A}^\top & \hat{C}^\top \\ \hat{B}^\top & \hat{D}^\top \end{bmatrix}^\top N_{V,W} \begin{bmatrix} I & 0 \\ 0 & I \\ \hat{A}^\top & \hat{C}^\top \\ \hat{B}^\top & \hat{D}^\top \end{bmatrix} \succcurlyeq 0 \right\},$$

where $N_{V,W}$ is given by (10) with $W := \mathrm{blkdiag}(\hat{W}, I_p)$ and $V := \mathrm{blkdiag}(\hat{V}, I_m)$.

Theorem 1 has a nice interpretation in terms of *data reduction*. Namely, the matrix $N_{V,W}$ characterizing all

$$N_{V,W} := \left[\begin{array}{c|c} W^\top(N|N_{22} + N_{12}N_{22}^{-1}V(V^\top N_{22}^{-1}V)^{-1}V^\top N_{22}^{-1}N_{12}^\top)W & W^\top N_{12}N_{22}^{-1}V(V^\top N_{22}^{-1}V)^{-1} \\ \hline (V^\top N_{22}^{-1}V)^{-1}V^\top N_{22}^{-1}N_{12}^\top W & (V^\top N_{22}^{-1}V)^{-1} \end{array}\right] \tag{10}$$

reduced-order models depends only on the projection matrices $\hat{V}, \hat{W}$ and the original data matrix N. As such, $N_{V,W}$ is constructed from the data and noise model only. Importantly, $N_{V,W}$ has a lower dimension than N and can thus be regarded as a reduced data matrix.

4. DATA-DRIVEN GENERALIZED BALANCED TRUNCATION

4.1 Data informativity for generalized Lyapunov balancing

Based on Section 2.2, one can introduce the following notion of informativity.

Definition 1. We say that the data (U_-, X, Y_-) are *informative for generalized Lyapunov balancing (GLB)* if there exist $P = P^\top > 0$ and $Q = Q^\top > 0$ such that

$$APA^\top - P + BB^\top < 0 \tag{11}$$

and

$$A^\top QA - Q + C^\top C < 0 \tag{12}$$

for all $(A, B, C, D) \in \mathbf{\Sigma}$.

From Definition 1, P and Q can be regarded as *common* generalized controllability and observability Gramian, respectively, for all systems explaining the data. We thus formalize the following problems.

Problem 1. Find necessary and sufficient conditions under which the data (U_-, X, Y_-) are informative for GLB. If the data are informative, then characterize the reduced-order models via data-driven balanced truncation and provide error bounds with respect to the true system.

Observe that the data are informative for GLB if and only if QMI (7) implies the existence of positive definite matrices P and Q such that (11) and (12) hold. Such QMI implications can be viewed as a generalization of the classical S-lemma (Yakubovich (1977)) and have been investigated in van Waarde et al. (2022a). Based on the results of van Waarde et al. (2022a) and van Waarde et al. (2022b), data informativity for GLB can be fully characterized in terms of feasibility of certain LMIs as stated next.

Theorem 2. Suppose that there exists $\bar{S}$ such that (9) holds. Define

$$N_{\mathcal{C}} := \begin{bmatrix} I_n & 0 \\ 0 & 0 \\ 0 & I_{n+m} \end{bmatrix}^\top N \begin{bmatrix} I_n & 0 \\ 0 & 0 \\ 0 & I_{n+m} \end{bmatrix}$$

and

$$N_{\mathcal{O}} := \begin{bmatrix} I_n & 0 \\ 0 & 0 \\ 0 & I_{n+p} \end{bmatrix}^\top N^\sharp \begin{bmatrix} I_n & 0 \\ 0 & 0 \\ 0 & I_{n+p} \end{bmatrix},$$

where

$$N^\sharp := \begin{bmatrix} 0 & -I_{n+m} \\ I_{n+p} & 0 \end{bmatrix} N^{-1} \begin{bmatrix} 0 & -I_{n+p} \\ I_{n+m} & 0 \end{bmatrix}.$$

Then, the data (U_-, X, Y_-) are informative for generalized Lyapunov balancing if and only if

(i) $\begin{bmatrix} X_- \\ U_- \end{bmatrix}$ has full row rank,

(ii) there exists $P = P^\top > 0$ and a scalar $\alpha > 0$ such that

$$\begin{bmatrix} P & 0 & 0 \\ 0 & -P & 0 \\ 0 & 0 & -I_m \end{bmatrix} - \alpha N_{\mathcal{C}} > 0, \tag{13}$$

(iii) there exists $Q = Q^\top > 0$ and a scalar $\beta > 0$ such that

$$\begin{bmatrix} Q & 0 & 0 \\ 0 & -Q & 0 \\ 0 & 0 & -I_p \end{bmatrix} - \beta N_{\mathcal{O}} > 0. \tag{14}$$

As a consequence, all systems in $\mathbf{\Sigma}$ are balanced by a common balancing transformation matrix T satisfying $TPT^\top = T^{-\top}QT^{-1} = \Sigma_H$ where Σ_H is a matrix of the form (3), containing the common GHSVs.

4.2 Reduced-order models

By applying the Petrov-Galerkin projection, the reduced-order models of all systems in $\mathbf{\Sigma}$ via generalized balanced truncation are contained in the set

$$\hat{\mathbf{\Sigma}} := \left\{ (\hat{W}^\top A\hat{V}, \hat{W}^\top B, C\hat{V}, D) : (A, B, C, D) \in \mathbf{\Sigma} \right\}$$

where $\hat{V} = T^{-1}\Pi$ and $\hat{W} = T^\top\Pi$ with Π given by $\Pi := \begin{bmatrix} I_r \\ 0 \end{bmatrix}$ and T is the common balancing transformation matrix.

Based on Theorem 1, we can characterize the set $\hat{\mathbf{\Sigma}}$ in terms of a quadratic matrix inequality.

Corollary 2. Suppose that there exists $\bar{S}$ such that (9) holds and the data (U_-, X, Y_-) are informative for generalized Lyapunov balancing with T the corresponding balancing transformation. Then,

$$\hat{\mathbf{\Sigma}} = \left\{ (\hat{A}, \hat{B}, \hat{C}, \hat{D}) : \begin{bmatrix} I & 0 \\ 0 & I \\ \hat{A}^\top & \hat{C}^\top \\ \hat{B}^\top & \hat{D}^\top \end{bmatrix}^\top N_{V,W} \begin{bmatrix} I & 0 \\ 0 & I \\ \hat{A}^\top & \hat{C}^\top \\ \hat{B}^\top & \hat{D}^\top \end{bmatrix} \geqslant 0 \right\},$$

where $N_{V,W}$ is given by (10) with $W = \text{blkdiag}(T^\top\Pi, I_p)$ and $V = \text{blkdiag}(T^{-1}\Pi, I_m)$.

Now, we can see that the set $\hat{\mathbf{\Sigma}}$ characterizes a *data reduction* by generalized balanced truncation. Namely, the reduced matrix $N_{V,W}$ depends only on the data matrix N and projection matrices $\hat{V} = T^{-1}\Pi$ and $\hat{W} = T^\top\Pi$, where now these projection matrices are also derived from the data only via Theorem 2.

From the definition of $\hat{\mathbf{\Sigma}}$ above, suppose that $(\hat{A}, \hat{B}, \hat{C}, \hat{D}) \in \hat{\mathbf{\Sigma}}$, then it is always a truncation of a model in $\mathbf{\Sigma}$. Therefore they satisfy

$$\|\Sigma - \hat{\Sigma}\|_{\mathcal{H}_\infty} \leqslant 2 \sum_{i=\ell+1}^{\kappa} \sigma_i \tag{15}$$

$$\left[\begin{array}{c|c} \text{blkdiag}(K_{11}, (\frac{1}{2} - \mu)I_p, -K_{11}, -\gamma^{-2}I_m) & \text{blkdiag}(K_{12}, -\mu I_p, -K_{12}, -\gamma^{-2}I_m) \\ \hline \text{blkdiag}(K_{12}^\top, -\mu I_p, -K_{12}^\top, -\gamma^{-2}I_m) & \text{blkdiag}(K_{22}, (\frac{1}{2} - \mu)I_p, -K_{22}, -\gamma^{-2}I_m) \end{array}\right] - \text{blkdiag}(\delta N, \eta N_{V,W}) > 0 \tag{16}$$

where Σ and $\hat{\Sigma}$ are systems whose realizations are in $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$, respectively, and σ_is are the neglected GHSVs. However, the bound (15) has little practical relevance as it characterizes the error between *one* reduced-order system in $\hat{\mathbf{\Sigma}}$ and its *corresponding* high-order system in $\mathbf{\Sigma}$.

Instead, recall that we are interested in a reduced-order approximation of the true system (4). This system is unknown, but is guaranteed to satisfy (5). As also the corresponding reduced-order system is unknown (but in $\hat{\mathbf{\Sigma}}$), a practical relevant error bound should hold for *any* selection of a high-order system (from $\mathbf{\Sigma}$) and *any* reduced-order system (from $\hat{\mathbf{\Sigma}}$).

4.3 Distance to the true system

Suppose that we consider a reduced-order system $\hat{\Sigma}_0$ of order $r < n$ with its realization $(\hat{A}_0, \hat{B}_0, \hat{C}_0, \hat{D}_0) \in \hat{\mathbf{\Sigma}}$. We will use $\hat{\Sigma}_0$ as an approximation of the true system Σ_{true}. To evaluate the quality of this approximation, note that

$$\|\hat{\Sigma}_0 - \Sigma_{\text{true}}\|_{\mathcal{H}_\infty} \leqslant \sup\left\{\|\hat{\Sigma} - \Sigma\|_{\mathcal{H}_\infty} : \Sigma \in \mathbf{\Sigma}, \hat{\Sigma} \in \hat{\mathbf{\Sigma}}\right\}. \quad (17)$$

Here, we have used the small abuse of notation $\Sigma \in \mathbf{\Sigma}$ to mean $(A, B, C, D) \in \mathbf{\Sigma}$, where (A, B, C, D) is a realization of Σ.

The computation of the bound on the right-hand side of (17) is stated in the following result.

Theorem 3. The bound $\|\hat{\Sigma} - \Sigma\|_{\mathcal{H}_\infty} < \gamma$ holds for any $\Sigma \in \mathbf{\Sigma}$ and any $\hat{\Sigma} \in \hat{\mathbf{\Sigma}}$ if and only if there exist a matrix $K = K^\top > 0$ in $\mathbb{R}^{(n+r)\times(n+r)}$ partitioned as

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{12}^\top & K_{22} \end{bmatrix}, \text{ with } K_{11} \in \mathbb{R}^{n\times n},$$

and scalars $\delta > 0$, $\eta > 0$ and μ such that (16) holds, where N and $N_{V,W}$ are given by (8) and (10), respectively.

In order to obtain the smallest upper bound, i.e., the smallest γ such that the conditions in Theorem 3 hold, one may solve (16) by minimizing γ using semidefinite programming (Nesterov and Nemirovskii, 1994, Sect. 6.4). It gives $\|\hat{\Sigma}_0 - \Sigma_{\text{true}}\| < \gamma$. Note that this upper bound is uniform for any $\hat{\Sigma}_0$ picked from $\hat{\mathbf{\Sigma}}$. Therefore, it can be regarded as an *a priori* error bound.

5. CONCLUSION

In this abstract, a data-driven procedure to obtain reduced-order models from noisy data is developed. The procedure begins with introducing the concept of data reduction. We then follow up the data reduction concept by constructing specific projection matrices from data. In particular, we provide necessary and sufficient conditions such that all systems explaining the data have common generalized Gramians. Subsequently, a common balancing transformation and therefore common projection matrices for generalized balanced truncation are available to apply the data reduction. As such, a set of reduced-order models via generalized balanced truncation can then be characterized in terms of a lower-dimensional QMI. Moreover, all reduced-order models in this set are guaranteed to be asymptotically stable and computable a priori upper bound on the reduction error with respect to the true system are available.

Some extensions of this work include exploiting data reduction via other projection-based reduction methods and using this approach by input-output data only. Moreover, investigating model reduction with preserving specific system properties such as network structure and port-Hamiltonian structure is often desirable.

REFERENCES

Antoulas, A.C. (2005). *Approximation of Large-Scale Dynamical Systems*. Advances in Design and Control, SIAM, Philadelphia, PA.

Beattie, C. and Gugercin, S. (2012). Realization-independent $\mathcal{H}_2$-approximation. In *Proceedings of the 51st IEEE Conference on Decision and Control*, 4953–4958. IEEE.

Burohman, A.M., Besselink, B., Scherpen, J.M.A., and Camlibel, M.K. (2021). From data to reduced-order models via generalized balanced truncation. *arXiv preprint arXiv:2109.11685*.

Burohman, A.M., Besselink, B., Scherpen, J.M.A., and Camlibel, M.K. (2020). From data to reduced-order models via moment matching. *arXiv preprint arXiv:2011.00150*.

Dullerud, G.E. and Paganini, F. (2000). *A Course in Robust Control Theory: A Convex Approach*, volume 36. Springer Science+Business Media, New York, NY.

Gosea, I.V., Gugercin, S., and Beattie, C. (2022). Data-driven balancing of linear dynamical systems. *SIAM Journal on Scientific Computing*, 44(1), A554–A582.

Markovsky, I., Willems, J.C., Rapisarda, P., and De Moor, B.L.M. (2005). Algorithms for deterministic balanced subspace identification. *Automatica*, 41(5), 755–766.

Nesterov, Y. and Nemirovskii, A. (1994). *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13. Studies in Applied Mathematics, SIAM, Philadelphia, PA.

Peherstorfer, B., Gugercin, S., and Willcox, K. (2017). Data-driven reduced model construction with time-domain Loewner models. *SIAM Journal on Scientific Computing*, 39(5), A2152–A2178.

Rapisarda, P. and Trentelman, H.L. (2011). Identification and data-driven model reduction of state-space representations of lossless and dissipative systems from noise-free data. *Automatica*, 47(8), 1721–1728.

Scarciotti, G. and Astolfi, A. (2017). Data-driven model reduction by moment matching for linear and nonlinear systems. *Automatica*, 79, 340–351.

van Waarde, H.J., Camlibel, M.K., and Mesbahi, M. (2022a). From noisy data to feedback controllers: Non-conservative design via a matrix S-lemma. *IEEE Transactions on Automatic Control*, 67(1), 162–175.

van Waarde, H.J., Eising, J., Trentelman, H.L., and Camlibel, M.K. (2020). Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11), 4753–4768.

van Waarde, H.J., Camlibel, M.K., Rapisarda, P., and Trentelman, H.L. (2022b). Data-driven dissipativity analysis: application of the matrix S-lemma. *IEEE Control Systems Magazine*, 42(3), 140–149.

Yakubovich, V.A. (1977). S-procedure in nonlinear control theory. *Vestnik Leningrad University Mathematics*, 4, 73–93.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Braess' paradox for power flow feasibility in DC power grids with constant-power loads ⋆,⋆⋆

Mark Jeeninga *

* *DISMA, Politecnico di Torino, Torino, Italy (e-mail: mark.jeeninga@polito.it)*

Abstract: This paper studies the power flow feasibility of DC power grids with constant-power loads. We introduce and motivate the concept of Braess' paradox for power flow feasibility, and show that this phenomenon can occur in most practical power grids with at least two source nodes.

Keywords: Power systems stability, Optimal operation and control of power systems, Smart grids

1. INTRODUCTION

Braess' paradox is a phenomenon in traffic flow networks, first observed in Braess (1968), described by the counterintuitive observation that adding a road (or improving the capacity of a road) does not necessarily improve the traffic flow. This phenomenon is a prevalent property of such networks (Steinberg and Zangwill (1983)). Similar observations for power grids are known. The papers Cohen and Horowitz (1991); Nagurney and Nagurney (2016) report several physical examples of two-port DC circuits for which the current flow exhibits behavior analogous to Braess' paradox in traffic flow networks. In these papers it is shown that the addition of a line can lead to an increase of the current flows in all lines, which goes against the intuition that adding a line allows for improved flow in the network, and is referred to as Braess' paradox. These phenomena have also been studied in Baillieul et al. (2015); Wang and Baillieul (2016), which considers the effects of adding a line (or an increase of the conductance of a line) in DC circuits with voltage-controlled or current-controlled nodes.

In this paper we study the power flow of DC power grids with fixed voltage sources and constant-power loads. Such loads may appear in practical power grids, and are known to destabilize the power grid due to their negative impedance characteristic (Emadi et al. (2006)). In these power grids it may occur that the sources cannot satisfy the power demands of the loads, in which case we say that the power flow is unfeasible. The feasibility of the power flow is important for their long-term operation, since sustained unfeasible power flow may lead to unintended behavior system such as voltage oscillations, voltage collapse, and blackouts (Kundur et al. (1994); Van Cutsem and Vournas (2008)). This feasibility problem is a classical problem in the literature (Tinney and Hart (1967); Hill and Mareels (1990); Löf et al. (1993)) and has gained more attention over the past decade (Bolognani and Zampieri (2015); Barabanov et al. (2016); Simpson-Porco et al. (2016); Matveev et al. (2020)). A full characterization of this feasibility problem has been presented in Jeeninga et al. (2020a,b).

The aim of the present paper is to study how the feasibility of the power flow in these systems is affected by changes in the line conductances. This study is primarily motivated by Witthaut and Timme (2012), where it was observed that increasing the conductance of a line in an AC power grid leads to voltage oscillations, which can be attributed to unfeasibility of the power flow after this increase. The contributions of the present paper are of a theoretical nature, and show that an analogue of Braess' paradox can also occur in DC power grids with constant-power loads, and that this may occur for most practical power grids with multiple sources.

The structure of this paper is as follows. In Section 2 we state the model for DC power grids with constant-power loads at steady state. In Section 3 we formulate the Braess' paradox of power flow feasibility, and show that Braess' paradox can occur in most practical power grids. Section 4 concludes the paper.

Notation

For a vector $x = (x_1 \ \cdots \ x_k)^\top$ we denote

$$[x] := \operatorname{diag}(x_1, \ldots, x_k).$$

We let $\mathbb{1}$ and $\mathbb{0}$ denote the all-ones and all-zeros vector, respectively, and let I denote the identity matrix. We let their dimensions follow from their context. We let e_i denote the i-th column of I. All vector and matrix inequalities are taken to be element-wise.

⋆ This submission is an extended abstract to the paper "Braess' paradox for power flow feasibility and parametric uncertainties in DC power grids with constant-power loads," Systems & Control Letters 161 (2022): 105146.

⋆⋆This work was partially supported by NWO (Dutch Research Council) project 'Energy management strategies for interconnected smart microgrids' within the DST-NWO Joint Research Program on Smart Grids, and by a MIUR grant Dipartimento di Eccellenza 2018–2022 [CUP: E11G18000350001]

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2. THE POWER GRID MODEL

Throughout this paper we study DC power grids at steady-state, and model such systems by a resistive circuit. We model a power grid consisting of n load nodes and m source nodes as follows. If distinct nodes i and j are connected by a line, we let $G_{ij} = G_{ji} > 0$ denote the conductance of this line. If the nodes are not connected by a line we put $G_{ij} = G_{ji} = 0$. The Kirchhoff matrix $Y \in \mathbb{R}^{(n+m)\times(n+m)}$ associated to the lines in the grid is defined by

$$Y_{ij} := \begin{cases} \sum_k G_{ki} & \text{if } i = j \\ -G_{ij} & \text{if } i \neq j \end{cases}. \tag{1}$$

The voltage potentials and injected currents at the loads are collected in the vectors $V \in \mathbb{R}^{n+m}$ and $\mathcal{I} \in \mathbb{R}^{n+m}$, respectively. The quantities V, $\mathcal{I}$ and Y are partioned as

$$V = \begin{pmatrix} V_L \\ V_S \end{pmatrix}; \quad \mathcal{I} = \begin{pmatrix} \mathcal{I}_L \\ \mathcal{I}_S \end{pmatrix}; \quad Y = \begin{pmatrix} Y_{LL} & Y_{LS} \\ Y_{SL} & Y_{SS} \end{pmatrix},$$

according to whether nodes are loads (L) or sources (S). We assume that the nodes in the power grid are connected, which means that $\mathbb{1}$ spans the kernel of Y, and that the principal submatrices Y_{LL} and Y_{SS} are positive definite. Due to Kirchhoff's and Ohm's laws we have $\mathcal{I} = YV$. We define the open-circuit voltages $V_L^* > \mathbb{0}$ to be the unique vector of voltage potentials at the loads such that $\mathcal{I}_L = \mathbb{0}$, which satisfies

$$V_L^* := -{Y_{LL}}^{-1} Y_{LS} V_S \tag{2}$$

(*e.g.*, see Van der Schaft (2010)). The power injected at the nodes is given by $P = [V]\mathcal{I}$. Since $\mathcal{I} = YV$ and due to (2) we have

$$P_L = [V_L](Y_{LL}V_L + Y_{LS}V_S) = [V_L]Y_{LL}(V_L - V_L^*) \tag{3a}$$
$$P_S = [V_S](Y_{SS}V_S + Y_{SL}V_L). \tag{3b}$$

The total power injected at the nodes is given by $\mathbb{1}^\top P = V^\top \mathcal{I}$.

We assume that all loads demand a constant power. Since we study the power grid at steady state, this is to say that we want to choose V_L such that all power demands are satisfied. The constant power demands are collected in the vector $P_\mathrm{c} \in \mathbb{R}^n$. The question if such a V_L exists for a given P_c gives rise to the DC power flow feasibility problem:

Definition 2.1. Given a power grid with Kirchhoff matrix Y, source voltages $V_S > \mathbb{0}$ and constant power demands P_c, we say that the power flow (of the power grid) is *feasible* if there exists a vector $V_L > \mathbb{0}$ of load voltages such that

$$P_\mathrm{c} = -P_L = [V_L]Y_{LL}(V_L^* - V_L).$$

Put differently, feasibility of a power grid means that the constant power demands at the loads can be satisfied at steady state. It is noted from (3a) that if Y_{LL} is reducible (or equivalently, block-diagonal), then power flow feasibility can be analyzed for each reducible component separately. Hence without loss of generality we assume that Y_{LL} is irreducible, which is equivalent to saying that the subgraph induced by the load nodes is connected.

The scalar $-\mathbb{1}^\top P_L$ represents the total amount of power that is drained by the loads. Intuitively, the total amount of power that can be drained by the loads is bounded from above. Recall that the power flow is feasible if we have that $P_\mathrm{c} = -P_L$. Hence this means that also $\mathbb{1}^\top P_\mathrm{c}$, the total power demand of the load nodes, is bounded from above whenever the power flow is feasible. The *maximizing power demand*, given by

$$P_\mathrm{max} := -\tfrac{1}{4}[V_L^*]Y_{LS}V_S, \tag{4}$$

is the unique vector of constant power demands for which the power flow is feasible and the total power demand is maximized:

Proposition 2.2. ((Jeeninga et al., 2020a, Lem. 2.17)). Consider a power grid with Y and $V_S > \mathbb{0}$ given. If $P_\mathrm{c} \in \mathbb{R}^n$ is a vector of power demands such that the power flow is feasible, then

$$\mathbb{1}^\top P_\mathrm{c} \leq \mathbb{1}^\top P_\mathrm{max}, \tag{5}$$

with equality if and only if $P_\mathrm{c} = P_\mathrm{max}$. Put differently, the quantity $\mathbb{1}^\top P_\mathrm{max}$ is the maximal total power demand that can be satisfied by the power grid. The unique voltage potentials corresponding to P_max are $V_L = \frac{1}{2}V_L^*$.

Note that (5) is a necessary condition for power flow feasibility.

3. BRAESS' PARADOX IN DC POWER GRIDS WITH CONSTANT-POWER LOADS

The classical formulation of Braess' paradox from Braess (1968); Cohen and Horowitz (1991); Nagurney and Nagurney (2016) is that, after increasing line conductances or adding a line in a DC circuit, the quality of the current flow in the power grid becomes measurably worse. In this paper however we follow a stronger formulation of Braess' paradox, inspired by Witthaut and Timme (2012), by studying how adding lines or increasing conductances can destabilize the power grid, and in particular, lose power flow feasibility. We formalize this as follows:

Definition 3.1. Braess' paradox for power flow feasibility is the phenomenon that adding a line or increasing a line conductance in a power grid destroys the feasibility of the power flow and destabilizes the power grid. We say that Braess' paradox for power flow feasibility can occur in a power grid if there exists a vector of power demands P_c such that the power flow is feasible and becomes unfeasible after increasing a line conductance or adding a line.

In the remainder of this paper we refer to Definition 3.1 simply as the Braess' paradox, for the sake of brevity.

3.1 A sufficient condition for the occurrence Braess' paradox

In this section we show how Braess' paradox may occur by studying the maximal total power demand $\mathbb{1}^\top P_\mathrm{max}$. It can be shown that an increase of a line conductance can make the maximal total power demand $\mathbb{1}^\top P_\mathrm{max}$ decrease. This means that the power flow in a power grid can become unfeasible after such an increase, in particular when the total power demand $\mathbb{1}^\top P_\mathrm{c}$ is close to its maximum $\mathbb{1}^\top P_\mathrm{max}$, such as when $P_\mathrm{c} = P_\mathrm{max}$. Hence, following Definition 3.1, we have that Braess' paradox occurs. We formalize this in the following theorem, which presents a sufficient condition for the existence of a pair of loads for which Braess' paradox occurs when the line between them is altered (or added).

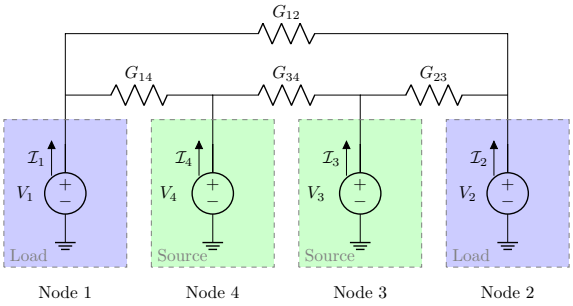


Fig. 1. A schematic depiction of a DC power grid with two loads and two sources ($n = 2$, $m = 2$) for which Braess' paradox can occur.

Theorem 3.2. Consider a power grid with Kirchhoff matrix Y, source voltages $V_S > \mathbb{0}$ and at least two loads and at least two sources. Let P_c be a vector of constant power demands such that the power flow is feasible. If there exist (distinct) load nodes i and j such that the open-circuit voltages (2) satisfy $(V_L^*)_i \neq (V_L^*)_j$, and if P_c satisfies

$$\mathbb{1}^\top P_c > \mathbb{1}^\top P_{\max} - \frac{(\frac{1}{2}(V_L^*)_i - \frac{1}{2}(V_L^*)_j)^2}{(e_i - e_j)^\top {Y_{LL}}^{-1}(e_i - e_j)}, \qquad (6)$$

then there exists a scalar $c > 0$ such that the power flow becomes unfeasible after increasing G_{ij} by c, either through increasing the conductance of the line between loads i and j, or adding a new line between loads i to j. In particular, if we consider $P_c = P_{\max}$, then the power flow becomes unfeasible for any increase of G_{ij}.

Theorem 3.2 tells us that if the open-circuit voltages are not all equal (*i.e.*, $V_L^* \notin \text{span}\{\mathbb{1}\}$), then Braess' paradox can occur when the power demands are close to the maximizing power demands $P_{\max}$. We illustrate this by the following example.

Example 3.3. Consider the power grid with two loads and two sources, as depicted in Figure 1, where $V_S = (1\ 3)^\top$, $G_{12} = 0.3$, $G_{14} = 1$, $G_{23} = 5$ and $G_{34} = 1$. The corresponding open-circuit voltages are $V_L^* = \frac{1}{6.8}(17.4\ 7.4)^\top$, and are not a multiple of $\mathbb{1}$. The blue area in Figure 2 depicts the set of all vectors P_c such that the power flow is feasible. Since $(V_L^*)_1 \neq (V_L^*)_2$, Theorem 3.2 states that we will observe Braess' paradox if we increase the line between load 1 and load 2. We increase the conductance of this line by 0.7. This results in the green area in Figure 2 corresponding to the set $\widehat{F}$ of P_c such that the power flow is feasible. It is observed that the blue area is not contained in the green area. Hence there are vectors of power demands for which the power flow has become unfeasible after increasing the conductance of the line. In particular we see that $P_{\max}$ is no longer feasible after the conductance is increased.

It should be noted that whenever only one source node is present, then all open-circuit voltages are equal and Theorem 3.2 cannot be applied. This is for example the case in distribution grids where all loads are constant-power loads, where the slack bus acts as the only source node in the grid.

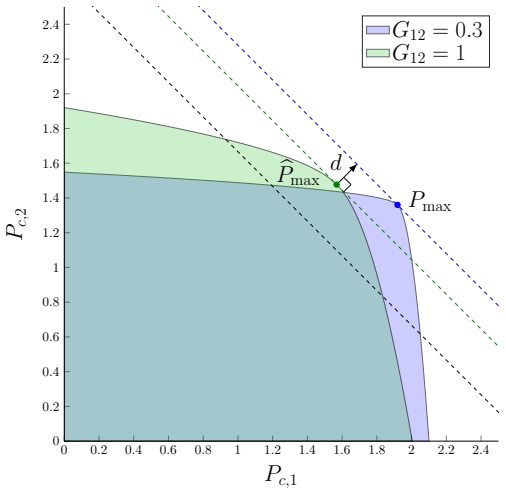


Fig. 2. Plots for the sets of constant power demands P_c for which the power flow of the power grid in Example 3.3 is feasible. Throughout this paper the plots of such sets have been obtained through the parametrization of the feasibility boundary presented in Jeeninga et al. (2020b). The blue and green regions correspond to respectively $G_{12} = 0.3$ and $G_{12} = 1$. This corresponds to increasing the conductance G_{12} of the line between node 1 and node 2 by 0.7 (see Figure 1). The blue and green dashed lines are the points for which equality holds in (5) for the respective power grids. The decrease of the maximal total power demand is indicated by d. We observe that $P_{\max}$ is no longer feasible after the conductance is increased. The black dashed line corresponds the points for which equality in (6) holds. Theorem 3.2 shows that Braess' paradox may occur for all feasible vectors of constant power demands beyond the black dashed line, such as for $P_{\max}$.

3.2 Prevalence of Braess' paradox for power flow feasibility

In Steinberg and Zangwill (1983) it was shown that Braess' paradox is a widespread phenomenon which can occur in most traffic flow networks. Similarly, Coletta and Jacquod (2016) showed that Braess' paradox may occur in any coupled-oscillator network such as AC power grids. In this section we validate that Braess' paradox is also common for DC power flow feasibility, and can occur in most practical DC power grids with constant-power loads, regardless of the grid topology.

It can be shown that the condition on the open-circuit voltages V_L^* may be restated in terms of the source voltages

Theorem 3.4. Consider a power grid with Y and $V_S > \mathbb{0}$ given and with at least two load nodes and at least two sources. If Y_{LS} has full column rank and the source voltages are not all equal (*i.e.*, $V_S \notin \text{span}\{\mathbb{1}\}$), then there exist a vector of power demands P_c and a line between two load nodes such that the power flow is feasible and becomes unfeasible after increasing the line conductance.

Theorem 3.4 shows that Braess' paradox may occur if Y_{LS} has full column rank and not all source voltages are equal. Apart from the condition on Y_{LS}, this statement does

not depend on the topology of the grid. Hence, Braess' paradox may occur in both radial (tree) and meshed grid topologies.

The condition that Y_{LS} has full column rank can be interpreted as the property that there are no redundant voltage sources, in the sense that the removal of one source cannot be compensated by other sources. A study on benchmark power grids was performed to illustrate that the condition that Y_{LS} has full-rank is prevalent in practical power grids. The results of this study are omitted for spatial considerations, but are available in the full version of the article.

For all power grids with a single source node we have that Y_{LS} has full column rank. This shows that the premises of Theorem 3.2 and Theorem 3.4 do not hold for power grids with a single source. However, since practical power grids commonly have multiple sources [1], and all source voltages are not likely to be the same, we conclude that Braess' paradox for power flow feasibility may occur in most practical DC power grids.

4. CONCLUSION AND DISCUSSION

In this paper we have shown that an analogue of Braess' paradox may occur in the power flow feasibility of a DC power grid. The observed phenomenon states that an increase of a line conductance has the potential to destroy the feasibility of the power flow in a power grid. We have shown that this phenomenon may occur in most practical power grids with multiple sources.

An interesting further direction of research is studying if Braess' paradox for power flow feasibility could also occur in the case where there is only one source node. Moreover, the full article of this abstract also studies what the implications are for the occurrence of Braess' paradox when only bounds of the line conductances are known. A sufficient condition is presented in that article, but the search for a necessary and sufficient condition is still open.

ACKNOWLEDGMENT

The author would like to thank Claudio De Persis and Arjan van der Schaft for their valuable suggestions and discussions.

REFERENCES

Baillieul, J., Zhang, B., and Wang, S. (2015). The Kirchhoff-Braess paradox and its implications for smart microgrids. In *2015 54th IEEE Conference on Decision and Control (CDC)*, 6556–6563.

Barabanov, N., Ortega, R., Griñó, R., and Polyak, B. (2016). On existence and stability of equilibria of linear time-invariant systems with constant power loads. *IEEE Trans. Circuits Syst. I, Reg. Papers*, 63(1), 114–121.

Bolognani, S. and Zampieri, S. (2015). On the existence and linear approximation of the power flow solution in power distribution networks. *IEEE Trans. Power Syst.*, 31(1), 163–172.

Braess, D. (1968). Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung*, 12(1), 258–268.

Cohen, J.E. and Horowitz, P. (1991). Paradoxical behaviour of mechanical and electrical networks. *Nature*, 352(6337), 699–701.

Coletta, T. and Jacquod, P. (2016). Linear stability and the Braess paradox in coupled-oscillator networks and electric power grids. *Phys. Rev. E*, 93, 032222.

Emadi, A., Khaligh, A., Rivetta, C.H., and Williamson, G.A. (2006). Constant power loads and negative impedance instability in automotive systems: definition, modeling, stability, and control of power electronic converters and motor drives. *IEEE Trans. Veh. Technol.*, 55(4), 1112–1125.

Hill, D.J. and Mareels, I.M.Y. (1990). Stability theory for differential/algebraic systems with application to power systems. *IEEE Trans. Circuits Syst.*, 37(11), 1416–1423.

Jeeninga, M., Persis, C.D., and van der Schaft, A.J. (2020a). DC power grids with constant-power loads—Part I: A full characterization of power flow feasibility, long-term voltage stability and their correspondence.

Jeeninga, M., Persis, C.D., and van der Schaft, A.J. (2020b). DC power grids with constant-power loads—Part II: nonnegative power demands, conditions for feasibility, and high-voltage solutions.

Kundur, P., Balu, N.J., and Lauby, M.G. (1994). *Power system stability and control*, volume 7. McGraw-hill New York.

Löf, P.A., Hill, D.J., Arnborg, S., and Andersson, G. (1993). On the analysis of long-term voltage stability. *International Journal of Electrical Power & Energy Systems*, 15(4), 229 – 237.

Matveev, A.S., Machado, J.E., Ortega, R., Schiffer, J., and Pyrkin, A. (2020). A tool for analysis of existence of equilibria and voltage stability in power systems with constant power loads. *IEEE Trans. Autom. Control*, 65(11), 4726–4740.

Nagurney, L.S. and Nagurney, A. (2016). Physical proof of the occurrence of the Braess paradox in electrical circuits. *EPL (Europhysics Letters)*, 115(2), 28004.

Simpson-Porco, J.W., Dörfler, F., and Bullo, F. (2016). Voltage collapse in complex power grids. *Nature Communications*, 7(10790).

Steinberg, R. and Zangwill, W. (1983). The prevalence of Braess' paradox. *Transportation Science*, 17.

Tinney, W.F. and Hart, C.E. (1967). Power flow solution by Newton's method. *IEEE Trans. Power App. Syst.*, PAS-86(11), 1449–1460.

Van Cutsem, T. and Vournas, C. (2008). *Voltage stability of electric power systems*. Springer Science & Business Media.

Van der Schaft, A. (2010). Characterization and partial synthesis of the behavior of resistive circuits at their terminals. *Systems & Control Letters*, 59(7), 423 – 428.

Wang, S. and Baillieul, J. (2016). Kirchhoff-Braess phenomena in DC electric networks. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 3286–3293.

Witthaut, D. and Timme, M. (2012). Braess's paradox in oscillator networks, desynchronization and power outage. *New J. Phys.*, 14(8), 083036.

[1] With the possible exception of power distribution grids with constant-power loads, which are commonly modeled with a single source or slack bus.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Guaranteeing a minimum distance to infeasibility in DC power grids with constant-power loads

Mark Jeeninga *

* *Politecnico di Torino, Corso Duca degli Abruzzi, 24 10129 Torino, Italy (e-mail: mark.jeeninga@polito.it)*

Abstract: This paper is concerned with the feasibility of the power flow in DC power power grids with constant power loads. Necessary and sufficient matrix inequalities are derived that guarantee a minimal p-norm distance between a configuration of power demands and the infeasibility boundary in the space of power demands. The (non)convexity of these matrix inequalities is studied subsequently.

Keywords: DC power grids, Power flow analysis, Matrix inequalities.

1. INTRODUCTION

Over the last decade, DC power grids have found an increasing interest among applications such as smart grids and high-voltage DC (HVDC) transmission. Currently, a major challenge in DC power grids is the presence of constant-power loads, which demand a constant amount of power from the grid. Such loads are known to destabilize the grid by selfishly extracting more power from the grid, which can lead to a collapse of nodal voltages known as voltage collapse. A particular example of such an event results from the case where the power demands cannot be met at steady state.

The power flow feasibility problem studies under which conditions constant-power demands can be satisfied at steady state. The problem has been fully characterized in Jeeninga et al. (2022a,b), resulting in necessary and sufficient conditions for power flow feasibility. Although these results are able to assess this feasibility problem, one shortcoming is that the robustness of power flow feasibility cannot be guaranteed. Such a guarantee is required when power grids should not be operated close to the feasibility boundary, or when the parameters of the system are uncertain. Sufficient conditions have been presented in Bolognani and Zampieri (2015) for a p-norm ball around the point where all power demands are zero, and a similar result concerning a polyhedral set was obtained in Simpson-Porco et al. (2016). To the best of the author's knowledge no other advances have been made. In the current paper, necessary and sufficient conditions are presented that guarantee the feasibility of the power demands within a p-norm ball around a given configuration of power demands.

Notation and matrix definitions

For a vector $x = (x_1 \ \cdots \ x_k)^\top$ we denote

$$[x] := \operatorname{diag}(x_1, \dots, x_k).$$

We let $\mathbb{1}$ and $\mathbb{0}$ denote the all-ones and all-zeros vector, respectively. We let their dimensions follow from their context. All vector and matrix inequalities are taken to be element-wise. We let e_i denote the i-th column of the identity matrix. We write $A \succeq B$ when $A - B$ is a symmetric positive semi-definite matrix, and $A \succneqq B$ when in addition $A \neq B$. We let $\|x\|_p$ denote the p-norm of $x \in \mathbb{R}^k$.

2. THE DC POWER GRID MODEL

This paper considers DC power grids with constant-power loads at steady state, which are modeled as a resistive circuit. Nodes (buses) in the grid are either sources (S) or loads (L). A source is a node at which the nodal voltage potentials of the network are fixed, such as a slack bus. A load is a node that demands a given quantity of power from the grid. The power flow feasibility problem asks if the nodal voltage potentials at the loads can be chosen such that all the power demands are satisfied.

To give a mathematical formulation of this problem we define the following quantities. We let $V = \begin{pmatrix} V_L^\top & V_S^\top \end{pmatrix}^\top \in \mathbb{R}^{n+m}$ be the voltage potentials at the nodes, which we assume to be positive. We let $Y \in \mathbb{R}^{(n+m)\times(n+m)}$ denote the Kirchhoff matrix of the power grid, which relates the voltage potentials in the grid to the nodal current $I \in \mathbb{R}^{n+m}$ injected into the network by $I = YV$. The power that is injected into the network at the loads is therefore given by

$$P_L = [V_L](Y_{LL}V_L + Y_{LS}V_S). \tag{1}$$

Let $P_c \in \mathbb{R}^n$ denote the power demands of the load nodes. The power flow feasibility problem asks if the demands P_c can be satisfied for some $V_L > \mathbb{0}$, in which case $P_L + P_c = \mathbb{0}$. We formalize this as follows.

Definition 2.1. Given Y and V_S, we say that P_c is feasible if there exists $V_L > \mathbb{0}$ such that

$$[V_L](Y_{LL}V_L + Y_{LS}V_S) + P_c = \mathbb{0}.$$

The set of all feasible P_c is denoted by $\mathcal{F}$.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2.1 Power flow feasibility as an LMI

It has been shown in Jeeninga et al. (2022b) that the set $\mathcal{F}$ is closed and convex, and that feasibility of P_c is equivalent to the feasibility of an LMI in terms of the matrix

$$Q_{P_c}(\lambda) := \begin{pmatrix} \frac{1}{2}([\lambda]Y_{LL} + Y_{LL}[\lambda]) & \frac{1}{2}[\lambda]Y_{LS}V_S \\ \frac{1}{2}([\lambda]Y_{LS}V_S)^\top & \lambda^\top P_c \end{pmatrix}.$$

We repeat the result for the sake of completeness.

Theorem 2.2. (LMI for power flow infeasibility). Given Y and V_S, the vector P_c is not feasible if and only if there exists a vector $\lambda > \mathbb{0}$ such that $Q_{P_c}(\lambda)$ is positive definite.

Theorem 2.2 tells us that power flow infeasibility is equivalent to the feasibility of an LMI. This equivalence may be rephrased by using the alternative of the LMI, *e.g.*. see Balakrishnan and Vandenberghe (2003). By doing so we obtain the following LMI condition which is equivalent to power flow feasibility.

Theorem 2.3. (LMI for power flow feasibility). Given Y and V_S, the vector P_c is feasible if and only if there exists a nonzero positive semi-definite matrix $Z = Z^\top \in \mathbb{R}^{(n+1)\times(n+1)}$ such that

$$\operatorname{trace}(Z\ Q_{P_c}(e_i)) \leq 0$$

for all $i = 1, \ldots, n$.

3. THE MINIMAL DISTANCE TO THE POWER FLOW FEASIBILITY BOUNDARY

In a practical setting it is desirable that a feasible P_c does not lie close to the feasibility boundary, for example in the case when only estimates of P_c are available. This note is therefore concerned with finding conditions that guarantee a minimal distance of P_c to the feasibility boundary $\partial\mathcal{F}$. Equivalently, we are interested in conditions that guarantee the feasibility of all vectors of power demands that lie in a neighborhood of P_c. To this end we let the p-norm ball centered at y be defined by

$$\mathcal{B}_{p,\gamma}(y) := \{\ z \mid \|y - z\|_p \leq \gamma\ \}, \quad \gamma > 0.$$

Additionally we would like to know if such conditions are computationally attractive, as in the case of Theorem 2.2 and Theorem 2.3. Our main problem is formalized as follows.

Problem 3.1. Let Y, V_S and P_c be given. Under what conditions do we have that any vector $\widehat{P}_c \in \mathcal{B}_{p,\gamma}(P_c)$ is feasible? Put differently, under what conditions does the inclusion $\mathcal{B}_{p,\gamma}(P_c) \subseteq \mathcal{F}$ hold? This is to say that the distance between a feasible P_c and the power flow feasibility boundary $\partial\mathcal{F}$ is at least γ.

This note answers Problem 3.1 by presenting matrix inequalities that are analogous to Theorem 2.2 and Theorem 2.3, and necessary and sufficient for the feasibility of a ball $\mathcal{B}_{p,\gamma}(P_c)$. In addition we prove that these matrix inequalities are (multiple) LMIs when $p = 1$ or $p = \infty$, and lead to non-convex matrix inequalities for other p.

3.1 The ∞-norm distance to $\partial\mathcal{F}$

For a given P_c, the ∞-norm ball $\mathcal{B}_{\infty,\gamma}(P_c)$ is the set of the vectors $\widehat{P}_c$ of the power demand such that

$$|P_{c,i} - \widehat{P}_{c,i}| \leq \gamma,$$

which is equivalent to the condition

$$P_c - \gamma\mathbb{1} \leq \widehat{P}_c \leq P_c + \gamma\mathbb{1}.$$

Note that the inequalities are element-wise. The following result from Jeeninga et al. (2022b) states that any vector that is element-wise dominated by a feasible vector of power demands is also feasible.

Lemma 3.2. If $y \in \mathcal{F}$ and $\widehat{y} \leq y$, then also $\widehat{y} \in \mathcal{F}$.

In particular Lemma 3.2 implies the equivalence

$$P_c + \gamma\mathbb{1} \in \mathcal{F} \ \Leftrightarrow\ \mathcal{B}_{\infty,\gamma}(P_c) \subseteq \mathcal{F}.$$

Consequently, to verify that all $\widehat{P}_c \in \mathcal{B}_{\infty,\gamma}(P_c)$ are feasible, it suffices to verify the feasibility of $P_c + \gamma\mathbb{1}$. Theorem 2.2 and Theorem 2.3 therefore imply the following characterization.

Theorem 3.3. Given Y and V_S and P_c, then some $\widehat{P}_c \in \mathcal{B}_{\infty,\gamma}(P_c)$ is not feasible if and only if there exists a vector $\lambda > \mathbb{0}$ such that $Q_{P_c+\gamma\mathbb{1}}(\lambda)$ is positive definite.

Theorem 3.4. Given Y and V_S and P_c, then all $\widehat{P}_c \in \mathcal{B}_{\infty,\gamma}(P_c)$ are feasible if and only if there exists a nonzero positive semi-definite matrix $Z = Z^\top \in \mathbb{R}^{(n+1)\times(n+1)}$ such that

$$\operatorname{trace}(Z\ Q_{P_c+\gamma\mathbb{1}}(e_i)) \leq 0$$

for all $i = 1, \ldots, n$.

In summary, we may check if the ∞-norm distance between a vector P_c and the power flow infeasibility boundary $\partial\mathcal{F}$ is at least γ by solving the LMI condition in either Theorem 3.3 or Theorem 3.4.

3.2 The 1-norm distance to $\partial\mathcal{F}$

For a given P_c, the 1-norm ball $\mathcal{B}_{1,\gamma}(P_c)$ is equivalent to the convex hull of the vectors

$$P_c - \gamma e_i, \quad P_c + \gamma e_i,$$

with $i = 1, \ldots, n$. Recall that the set $\mathcal{F}$ is convex, and we therefore have that

$$\mathcal{B}_{1,\gamma}(P_c) \subseteq \mathcal{F} \ \Leftrightarrow\ P_c \pm \gamma e_i \in \mathcal{F} \tag{2}$$

for all $i = 1, \ldots, n$. Moreover, since $P_c - \gamma e_i \leq P_c + \gamma e_i$, Lemma 3.2 implies that it suffices to take only the positive signs in (2). By virtue to Theorem 2.2 and Theorem 2.3 we obtain the following characterization.

Theorem 3.5. Given Y and V_S and P_c, then some $\widehat{P}_c \in \mathcal{B}_{1,\gamma}(P_c)$ is not feasible if and only if there exists a vector $\lambda > \mathbb{0}$ such that $Q_{P_c+\gamma e_i}(\lambda)$ is positive definite for some $i \in \{1, \ldots, n\}$.

Theorem 3.6. Given Y and V_S and P_c, then all $\widehat{P}_c \in \mathcal{B}_{1,\gamma}(P_c)$ are feasible if and only if for each $i \in \{1, \ldots, n\}$ there exists a nonzero positive semi-definite matrix $Z_i = Z_i^\top \in \mathbb{R}^{(n+1)\times(n+1)}$ such that

$$\operatorname{trace}(Z_i\ Q_{P_c+\gamma e_i}(e_j)) \leq 0$$

for all $j = 1, \ldots, n$.

Note that Theorem 3.5 or Theorem 3.6 are equivalent to solving n separate LMIs. To summarize, we may check if the 1-norm distance between a vector P_c and the power flow infeasibility boundary $\partial\mathcal{F}$ is at least γ by solving the n LMI conditions in either Theorem 3.5 or Theorem 3.6.

3.3 The p-norm distance to $\partial\mathcal{F}$

In the remainder of this section we assume that p lies in the open interval $(1,\infty)$ and we let $q := \frac{p}{p-1}$ such that

$$\tfrac{1}{p} + \tfrac{1}{q} = 1.$$

For a given P_c, the p-norm ball $\mathcal{B}_{p,\gamma}(P_\mathrm{c})$ is the set of the vectors $\widehat{P}_\mathrm{c}$ of the power demand such that

$$\|P_\mathrm{c} - \widehat{P}_\mathrm{c}\|_p \leq \gamma. \tag{3}$$

We define $\delta = \widehat{P}_\mathrm{c} - P_\mathrm{c}$ for the sake of brevity. By Hölder's inequality (*e.g.*, see Roman (2008)) we know that

$$\sum_{i=1}^{n} |\lambda_i \delta_i| \leq \|\delta\|_p \|\lambda\|_q. \tag{4}$$

Moreover, equality holds in (4) for some λ_δ that lies in the same quadrant as δ, in which case the left-hand side of (4) equals ${\lambda_\delta}^\top \delta$. Using (3) we therefore have the inequality

$${\lambda_\delta}^\top \delta \leq \gamma \|\lambda_\delta\|_q.$$

Finally, by recalling that $\delta = \widehat{P}_\mathrm{c} - P_\mathrm{c}$ we conclude that

$${\lambda_\delta}^\top \widehat{P}_\mathrm{c} = {\lambda_\delta}^\top P_\mathrm{c} + {\lambda_\delta}^\top \delta \leq {\lambda_\delta}^\top P_\mathrm{c} + \gamma \|\lambda_\delta\|_q. \tag{5}$$

In the context of Theorem 2.2, inequality (5) implies that if $\widehat{P}_\mathrm{c}$ is not feasible then there exists a $\lambda > \mathbb{0}$ such that

$$Q_{P_\mathrm{c},q,\gamma}(\lambda) := \begin{pmatrix} \frac{1}{2}([\lambda]Y_{LL} + Y_{LL}[\lambda]) & \frac{1}{2}[\lambda]Y_{LS}V_S \\ \frac{1}{2}([\lambda]Y_{LS}V_S)^\top & \lambda^\top P_\mathrm{c} + \gamma\|\lambda\|_q \end{pmatrix}.$$

is positive definite. In fact, by considering all $\widehat{P}_\mathrm{c}$ in the ball $\mathcal{B}_{p,\gamma}(P_\mathrm{c})$ we may prove the following equivalence condition.

Theorem 3.7. Given Y and V_S and P_c, then some $\widehat{P}_\mathrm{c} \in \mathcal{B}_{p,\gamma}(P_\mathrm{c})$ is not feasible if and only if there exists a vector $\lambda > \mathbb{0}$ such that $Q_{P_\mathrm{c},q,\gamma}(\lambda)$ is positive definite, where $\frac{1}{p} + \frac{1}{q} = 1$.

Note that the above argumentation is not complete, and that a full proof is omitted for the same of brevity. An analogous condition to Theorem 2.3 may also be obtained by virtue of Hölder's inequality.

Theorem 3.8. Given Y and V_S and P_c, then all $\widehat{P}_\mathrm{c} \in \mathcal{B}_{p,\gamma}(P_\mathrm{c})$ are feasible if and only if there exists a nonzero positive semi-definite matrix $Z = Z^\top \in \mathbb{R}^{(n+1)\times(n+1)}$ such that

$$\operatorname{trace}\left(Z\ Q_{P_\mathrm{c},q,\gamma}(\lambda)\right) \leq 0$$

for all $\lambda \geq \mathbb{0}$, where $\frac{1}{p} + \frac{1}{q} = 1$.

To conclude we section we will analyze the (non)convexity of the matrix inequalities in Theorem 2.2 and Theorem 2.3. The matrix inequalities in Theorem 3.7 and Theorem 3.8 are not LMIs since the map $\lambda \mapsto Q_{P_\mathrm{c},q,\gamma}(\lambda)$ is not linear. Linear matrix inequalities are computationally attractive since they describe a convex problem, which have the property that local optima are also global optima. Moreover, algorithms to find these optima are well-studied.

Focusing on Theorem 3.7, the corresponding matrix inequality would be convex if the map $\lambda \mapsto Q_{P_\mathrm{c},q,\gamma}(\lambda)$ would have to satisfy

$$\begin{aligned}\alpha Q_{P_\mathrm{c},q,\gamma}(\lambda_1) + (1-\alpha) Q_{P_\mathrm{c},q,\gamma}(\lambda_2) \\ \preceq Q_{P_\mathrm{c},q,\gamma}(\alpha\lambda_1 + (1-\alpha)\lambda_2)\end{aligned} \tag{6}$$

for $\alpha \in (0,1)$, where we let λ_1, λ_2 such that $\mathbb{1}^\top \lambda_i = 1$. This relationship implies the existence of a $\widehat{\lambda}$ that satisfies $\mathbb{1}^\top \widehat{\lambda} = 1$ and for which $Q_{P_\mathrm{c},q,\gamma}(\widehat{\lambda})$ is the unique maximal element of the partial ordering induced by $\preceq$. Consequently, the matrix inequality in Theorem 3.7 may be evaluated by finding this $\widehat{\lambda}$ and checking if $Q_{P_\mathrm{c},q,\gamma}(\widehat{\lambda})$ is positive definite. However, we will show that the inequality (6) only holds when $\lambda_1 = \lambda_2$.

The triangle inequality implies that for vectors λ_1, λ_2 such that $\mathbb{1}^\top \lambda_i = 1$ we have

$$\alpha\|\lambda_1\|_q + (1-\alpha)\|\lambda_2\|_q \geq \|\alpha\lambda_1 + (1-\alpha)\lambda_2\|_q \tag{7}$$

for $\alpha \in (0,1)$. Moreover, since $q \in (1,\infty)$ we have that if $\lambda_1 \neq \lambda_2$ then the inequality in (7) is strict (*e.g.*, see Chapter 11 of Carothers (2004)). For such $\lambda_1 \neq \lambda_2$ we therefore have

$$\begin{aligned}\alpha Q_{P_\mathrm{c},q,\gamma}(\lambda_1) + (1-\alpha) Q_{P_\mathrm{c},q,\gamma}(\lambda_2) \\ \succneqq Q_{P_\mathrm{c},q,\gamma}(\alpha\lambda_1 + (1-\alpha)\lambda_2)\end{aligned}$$

for $\alpha \in (0,1)$, which implies that the matrix inequality is strictly concave, and that (6) only holds when $\lambda_1 = \lambda_2$.

By a similar argumentation it can be shown that Theorem 3.8 is a non-convex problem.

In summary, for $p \in (1,\infty)$ we may check if the p-norm distance between a vector P_c and the power flow infeasibility boundary $\partial\mathcal{F}$ is at least γ by solving the non-convex matrix inequality in either Theorem 3.7 or Theorem 3.8.

Although the matrix inequality in Theorem 3.7 is not convex, it is remarked that in special cases a unique maximizer of the partial ordering induced by $\preceq$ may exist. If this is the case, an algorithm that obtains a local maximizer will therefore also obtain the global maximizer.

4. CONCLUSION

This note studied the distance of a feasible vector of power demands to the power flow feasibility boundary. A p-norm ball around a given vector of power demands was considered, and it was shown that to verify that all vectors of power demands that lie within the ball are feasible, one may solve one of two matrix inequalities. In the case of $p = \infty$ this gives rise to a single LMI, in the case of $p = 1$, this gives rise to n LMIs, and in the case of $p \in (1,\infty)$ this gives rise to a non-convex (strictly concave) matrix inequality.

REFERENCES

Balakrishnan, V. and Vandenberghe, L. (2003). Semidefinite programming duality and linear time-invariant systems. *IEEE Transactions on Automatic Control*, 48(1), 30–41.

Bolognani, S. and Zampieri, S. (2015). On the existence and linear approximation of the power flow solution in power distribution networks. *IEEE Transactions on Power Systems*, 31(1), 163–172.

Carothers, N. (2004). *A short course on Banach space theory*, volume 64. Cambridge University Press.

Jeeninga, M., Persis, C.D., and van der Schaft, A.J. (2022a). DC power grids with constant-power loads—Part I: A full characterization of power flow feasibility, long-term voltage stability and their correspondence. *EEE Trans. Autom. Cont*, to appear.

Jeeninga, M., Persis, C.D., and van der Schaft, A.J. (2022b). DC power grids with constant-power loads—Part II: nonnegative power demands, conditions for feasibility, and high-voltage solutions. *EEE Trans. Autom. Cont*, to appear.

Roman, S. (2008). *Advanced linear algebra*, volume 3. Springer.

Simpson-Porco, J.W., Dörfler, F., and Bullo, F. (2016). Voltage collapse in complex power grids. *Nature Communications*, 7(10790). doi:10.1038/ncomms10790.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On a solution of the multidimensional truncated matrix-valued moment problem

David P. Kimsey * **Matina Trachana** **

* *Newcastle University, School of Mathematics, Statistics and Physics Newcastle upon Tyne, NE1 7RU UK (e-mail: david.kimsey@newcastle.ac.uk).*
** *Cardiff University, School of Mathematics, Cardiff, Wales CF24 4AG UK (e-mail: tranchanam@cardiff.ac.uk)*

Abstract: We will consider the multidimensional truncated $p \times p$ Hermitian matrix-valued moment problem. We will prove a characterisation of truncated $p \times p$ Hermitian matrix-valued multisequence with a minimal positive semidefinite matrix-valued representing measure via the existence of a flat extension, i.e., a rank preserving extension of a multivariate Hankel matrix (built from the given truncated matrix-valued multisequence). Moreover, the support of the representing measure can be computed via the intersecting zeros of the determinants of matrix-valued polynomials which describe the flat extension. We will also use a matricial generalisation of Tchakaloff's theorem due to the first author together with the above result to prove a characterisation of truncated matrix-valued multisequences which have a representing measure. When $p = 1$, our result recovers the celebrated flat extension theorem of Curto and Fialkow. The bivariate quadratic matrix-valued problem and the bivariate cubic matrix-valued problem are explored in detail.

Keywords: Sum-of-squares, semigroup and operator theory (47A57)

We will investigate the *multidimensional truncated matrix-valued moment problem*. Given a truncated multisequence $S = (S_\gamma)_{\substack{0 \le |\gamma| \le m \\ \gamma \in \mathbb{N}_0^d}}$, where $S_\gamma \in \mathcal{H}_p$ (i.e., S_γ is a $p \times p$ Hermitian matrix), we wish to find necessary and sufficient conditions on S for the existence of a $p \times p$ positive matrix-valued measure T on $\mathbb{R}^d$, with convergent moments, such that

$$S_\gamma = \int_{\mathbb{R}^d} x^\gamma dT(x) := \int \cdots \int_{\mathbb{R}^d} \prod_{j=1}^d x_j^{\gamma_j} dT(x_1, \ldots, x_d) \tag{1}$$

for all $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}_0^d$ such that $0 \le |\gamma| \le m$. We would also like to find a positive matrix-valued measure $T = \sum_{a=1}^\kappa Q_a \delta_{w^{(a)}}$ on $\mathbb{R}^d$ such that (1) holds and and

$$\sum_{a=1}^\kappa \operatorname{rank} Q_a \text{ is as small as possible,} \tag{2}$$

i.e., T is a finitely atomic measure of the form

$$T = \sum_{a=1}^\kappa \delta_{w^{(a)}} Q_a$$

with

$$\sum_{a=1}^\kappa \operatorname{rank} Q_a = \operatorname{rank} M(n).$$

If (1) holds, then T is called a *representing measure* for S. If (1) and (2) are in force, then T is called a *minimal representing measure for* S.

Before proceeding any further, we will first introduce frequently used notation. Commonly used sets are $\mathbb{N}_0, \mathbb{R}, \mathbb{C}$ denoting the sets of nonnegative integers, real numbers and complex numbers respectively. Given a nonempty set E, we let

$$E^d = \{(x_1, \ldots, x_d) : x_j \in E \text{ for } j = 1, \ldots, d\}.$$

Next, we let $\mathbb{C}^{p \times p}$ denote the set of $p \times p$ matrices with entries in $\mathbb{C}$ and $\mathcal{H}_p \subseteq \mathbb{C}^{p \times p}$ denote the set of $p \times p$ Hermitian matrices with entries in $\mathbb{C}$. Given

$$x = (x_1, \ldots, x_d) \in \mathbb{R}^d \quad \text{and} \quad \lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{N}_0^d,$$

we define

$$x^\lambda = \prod_{j=1}^d x_j^{\lambda_j} \quad \text{and} \quad |\lambda| = \lambda_1 + \cdots + \lambda_d$$

and

$$\Gamma_{m,d} := \{\gamma \in \mathbb{N}_0^d : 0 \le |\gamma| \le m\}.$$

We shall let $\mathbb{C}^{p \times p}[x_1, \ldots x_d]$ denote the set of *matrix polynomials* in the indeterminates $x_1, \ldots, x_d$, i.e., the set of all polynomials $P(x) = \sum_{\lambda \in \Gamma_{n,d}} x^\lambda P_\lambda$, where $P_\lambda \in \mathbb{C}^{p \times p}$ and n is arbitrary. We will also let

$$\mathbb{C}_n^{p \times p}[x_1, \ldots, x_d] := \{P \in \mathbb{C}^{p \times p}[x_1, \ldots, x_d] : \text{the total degree of } P(x) \le n\}.$$

In order to communicate our main contributions, we will need the notion of a d-Hankel matrix. Let $S := (S_\gamma)_{\gamma \in \Gamma_{2n,d}}$ be a given truncated $\mathcal{H}_p$-valued multisequence and $M(n)$ be the corresponding *d-Hankel matrix* based on S and defined as follows. We label the block rows and block columns by a family of monomials $(x^\gamma)_{\gamma \in \Gamma_{n,d}}$ ordered by the grader lexicographic ordering $\prec_{\text{grlex}}$. We let the entry in the block row indexed by x^γ and in the block column indexed by $x^{\tilde{\gamma}}$ be given by

$$S_{\gamma + \tilde{\gamma}}.$$

For example if $d = n = 2$, then

$$M(2) = \begin{array}{c} \\ 1 \\ X \\ Y \\ X^2 \\ XY \\ Y^2 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & X & Y & X^2 & XY & Y^2 \end{array} \\ \begin{pmatrix} S_{00} & S_{10} & S_{01} & S_{20} & S_{11} & S_{02} \\ S_{10} & S_{20} & S_{11} & S_{30} & S_{21} & S_{12} \\ S_{01} & S_{11} & S_{02} & S_{21} & S_{12} & S_{03} \\ S_{20} & S_{30} & S_{21} & S_{40} & S_{31} & S_{22} \\ S_{11} & S_{21} & S_{12} & S_{31} & S_{22} & S_{13} \\ S_{02} & S_{12} & S_{03} & S_{22} & S_{13} & S_{04} \end{pmatrix} \end{array}.$$

If we are given $S := (S_\gamma)_{\gamma \in \mathbb{N}_0^d}$, then shall let $M(\infty)$ denote the *infinite d-Hankel matrix* based on S, which can be defined analogously by letting the block rows and columns be indexed by $(x^\lambda)_{\lambda \in \mathbb{N}_0^d}$ (ordered by $\preceq_{\text{grlex}}$). If $d = 2$, then

$$M(\infty) = \begin{array}{c} \\ 1 \\ X \\ Y \\ X^2 \\ XY \\ Y^2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & X & Y & X^2 & XY & Y^2 & \cdots \end{array} \\ \begin{pmatrix} S_{00} & S_{10} & S_{01} & S_{20} & S_{11} & S_{02} & \cdots \\ S_{10} & S_{20} & S_{11} & S_{30} & S_{21} & S_{12} & \cdots \\ S_{01} & S_{11} & S_{02} & S_{21} & S_{12} & S_{03} & \cdots \\ S_{20} & S_{30} & S_{21} & S_{40} & S_{31} & S_{22} & \cdots \\ S_{11} & S_{21} & S_{12} & S_{31} & S_{22} & S_{13} & \cdots \\ S_{02} & S_{12} & S_{03} & S_{22} & S_{13} & S_{04} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}.$$

Let $S := (S_\gamma)_{\gamma \in \Gamma_{2n,d}}$ be a truncated $\mathcal{H}_p$-valued multisequence and let $M(n)$ be the corresponding d-Hankel matrix. Corresponding to any $P(x) = \sum_{\lambda \in \Gamma_{n,d}} x^\lambda P_\lambda \in \mathbb{C}_n^{p \times p}[x_1, \ldots, x_d]$, let $P(X)$ denote the element of $\mathbb{C}^{\binom{(n+d)!}{n!\,d!} p \times p}$ arising from $\sum_{\lambda \in \Gamma_{n,d}} x^\lambda P_\lambda$ when we replace x^λ by the corresponding block column in $M(n)$. The *variety of* $M(n)$, denoted by $\mathcal{V}(M(n))$, is given by

$$\mathcal{V}(M(n)) := \bigcap_{\substack{P \in \mathbb{C}_n^{p \times p}[x_1, \ldots, x_d] \\ P(X) = \operatorname{col}(0_{p \times p})_{\gamma \in \Gamma_{n,d}}}} \mathcal{Z}(\det(P(x))).$$

Main contributions

(C1) We will characterise positive infinite d-Hankel matrices with finite rank based on a $\mathcal{H}_p$-valued multisequence via an integral representation. Indeed, we will see that $S^{(\infty)} = (S_\gamma)_{\gamma \in \mathbb{N}_0^d}$ gives rise to a positive infinite d-Hankel matrix $M(\infty)$ with finite rank if and only if there exists a finitely atomic positive $\mathcal{H}_p$-valued measure T on $\mathbb{R}^d$ such that

$$S_\gamma = \int_{\mathbb{R}^d} x^\gamma \, dT(x) \qquad \text{for} \quad \gamma \in \mathbb{N}_0^d.$$

In this case, the support of the positive $\mathcal{H}_p$-valued measure T agrees with

$$\mathcal{V}(\mathcal{I}) := \bigcap_{P \in \mathcal{I} \subseteq \mathbb{C}^{p \times p}[x_1, \ldots, x_d]} \mathcal{Z}(\det P(x)),$$

where $\mathcal{V}(\mathcal{I})$ is the variety of a right ideal of matrix-valued polynomials based on the kernel of $M(\infty)$ and the cardinality of the support of T is exactly $\operatorname{rank} M(\infty)$.

(C2) Let $S = (S_\gamma)_{\gamma \in \Gamma_{2n,d}}$ be a given truncated $\mathcal{H}_p$-valued multisequence. We will see that S has a minimal representing measure $T = \sum_{a=1}^{\kappa} Q_a \delta_{w^{(a)}}$, i.e., $\sum_{a=1}^{\kappa} \operatorname{rank} Q_a = \operatorname{rank} M(n)$, if and only if the corresponding d-Hankel matrix $M(n)$ based on S has a *flat extension* $M(n+1)$, i.e., a positive rank preserving extension. In this case, the support of T agrees with $\mathcal{V}(M(n+1))$, where $\mathcal{V}(M(n+1))$ is the variety of the d-Hankel matrix $M(n+1)$ and $\sum_{a=1}^{\kappa} \operatorname{rank} Q_a = \operatorname{rank} M(n)$.

(C3) Let S be as in (C2). S has a representing measure if and only if the corresponding d-Hankel matrix $M(n)$ have an eventual extension $M(n+k)$ which admits a flat extension.

(C4) Let $S = (S_{00}, S_{10}, S_{01}, S_{20}, S_{11}, S_{02})$ be a given $\mathcal{H}_p$-valued truncated bisequence. We will see that necessary and sufficient conditions for S to have a minimal representing measure consist of $M(1)$ being positive semidefinite and a system of matrix equations having a solution. More precisely, if $M(1) \succeq 0$ and there exist $S_{30}, S_{21}, S_{12}, S_{03} \in \mathcal{H}_p$ such that

$$\operatorname{Ran} \begin{pmatrix} S_{20} & S_{11} & S_{02} \\ S_{30} & S_{21} & S_{12} \\ S_{21} & S_{12} & S_{03} \end{pmatrix} \subseteq \operatorname{Ran} M(1)$$

(hence, there exists $W = (W_{ab})_{a,b=1}^3 \in \mathbb{C}^{3p \times 3p}$ such that $M(1)W = B$, where

$$B = \begin{pmatrix} S_{20} & S_{11} & S_{02} \\ S_{30} & S_{21} & S_{12} \\ S_{21} & S_{12} & S_{03} \end{pmatrix})$$

and moreover, the following matrix equations hold:

$$W_{11}^* S_{11} + W_{21}^* S_{21} + W_{31}^* S_{12} = S_{11} W_{11} + S_{21} W_{21} + S_{12} W_{31}, \tag{3}$$

$$W_{13}^* S_{20} + W_{23}^* S_{30} + W_{33}^* S_{21} = W_{12}^* S_{11} + W_{22}^* S_{21} + W_{32}^* S_{12} \tag{4}$$

and

$$W_{12}^* S_{02} + W_{22}^* S_{12} + W_{32}^* S_{03} = S_{02} W_{12} + S_{12} W_{22} + S_{03} W_{32}, \tag{5}$$

then S has a minimal representing measure.

We will also see that if $M(1)$ is positive definite and obeys an extra condition (which is automatically satisfied if $p = 1$), then S has a minimal representing measure. However, if $M(1)$ is singular and $p \geq 2$, then S need not have a minimal representing measure.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The Truncated Moment Problem for Unital Commutative R–Algebras

Raúl E. Curto [*] **Mehdi Ghasemi** [**] **Maria Infusino** [***]
Salma Kuhlmann [****]

[*] *Department of Mathematics, University of Iowa, Iowa City, IA 52242, USA (e-mail: raul-curto@uiowa.edu).*
[**] *Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, S7N 5E6, Canada (e-mail: mehdi.ghasemi@usask.ca)*
[***] *Dipartimento di Matematica e Informatica, Universitá degli Studi di Cagliari, Palazzo delle Scienze, Via Ospedale 72, 09124 Cagliari, Italy (e-mail: maria.infusino@unica.it)*
[****] *Fachbereich Mathematik und Statistik, Universität Konstanz, Universitätstrasse 10, 78457 Konstanz, Germany (e-mail: salma.kuhlmann@uni-konstanz.de*

Keywords: truncated moment problem, full moment problem, measure, integral representation, linear functional.
2020 *Mathematics Subject Classification.* Primary: 44A60, 47A57, 28C05.

Extended Abstract

The Classical Truncated Moment Problem (TMP) dates back to the start of the twentieth century, and was initially developed by a number of mathematicians, including A.A. Markov, H. Hamburger, N.I. Akhiezer, M.G. Krein, A.A. Nudel'man, M. Riesz and I.S. Iohvidov. The theory ran parallel to the developments in the full moment problem, where the main focus was placed. Many decades later, renewed interest in TMP arose in connection with the so-called Subnormal Completion Problem (SCP) for unilateral weighted shifts. In 1966, J. Stampfli (52) proved that for any three positive numbers $a < b < c$, it is always possible to build a unilateral weighted shift W_α acting on $\ell^2(\mathbf{N}_0)$, with $\alpha \in \ell^\infty(\mathbf{N}_0)$, having initial weights $\alpha_0 = a$, $\alpha_1 = b$, $\alpha_2 = c$, and such that W_α is subnormal. In (9; 10), R.E. Curto and L.A. Fialkow solved the SCP for unilateral weighted shifts, by finding necessary and sufficient conditions for a finite collection of positive numbers to be the initial segment of weights of a subnormal unilateral weighted shift. Their approach was based on the fact that subnormality is detected by the existence of a positive Radon measure on the closed interval $[0, \|W_\alpha\|^2]$ whose moments are the moments γ_k of the weight sequence α, defined recursively as $\gamma_0 := 1$ and $\gamma_{k+1} := \alpha_k^2 \gamma_k$ (for all $k \in \mathbf{N}_0$). Thus, the subnormality of W_α is intrinsically related to a TMP. In the process, Curto and Fialkow proved the so-called Flat Extension Theorem for moment matrices, which is an essential component of their TMP theory in one and several real or complex variables.

A few years after the Curto-Fialkow results were published, J.B. Lasserre discovered some significant connections between real algebraic geometry, moment problems and polynomial optimization; he introduced a method known as semidefinite relaxations (see, e.g., (34)), which led to renewed interest in solutions of TMP, especially those with finitely atomic representing measures. The importance of polynomial optimization problems and the convenience of working with polynomials as algebraic and computational objects as well as intensive research on this area, is one of the main motivations for the study of moment problems for the algebra of polynomials. For ample information on the above mentioned developments, the reader is referred to (11; 12; 13; 14; 15; 16; 17; 18; 23; 24; 29; 30; 31; 34; 35; 36; 37; 38; 43; 47; 48; 50; 54; 57).

A fundamental tool in all those works is positivity. Given a closed subset K of ^{n}, a linear functional L defined on a subspace of $[\underline{X}] := [X_1, \ldots, X_n]$ is said to be $K-$*positive* when it assumes nonnegative values in all the elements of its domain which are nonnegative on K. For a set $\mathcal{A}$ of monomials in $[\underline{X}]$, a closed subset K of ^{n}, and a $K-$positive linear functional L on the Span($\mathcal{A}$), the $\mathcal{A}$–truncated K–moment problem is the question of establishing whether L can be represented as an integral with respect to a positive Radon measure whose support is contained in K. If such a measure exists then it is called a K–representing measure for L. The $\mathcal{A}-$Truncated $K-$Moment Problem terminology was introduced by J. Nie in (45), although he only considered the case when the set $\mathcal{A}$ is finite. When $\mathcal{A} = \{\underline{X}^\alpha : \alpha \in \mathbf{N}_0^n, |\alpha| \leq d\}$, for some $d \in \mathbf{N}$, the $\mathcal{A}$–truncated K–moment problem is usually known as the Classical K–TMP.

The Full K–Moment Problem, for closed $K \subseteq ^n$, corresponds to the case when $\mathcal{A} = \{\underline{X}^\alpha : \alpha \in \mathbf{N}_0^n\}$; that is, given a $K-$positive linear functional L on $[\underline{X}]$, find a criterion for the existence of a positive Radon measure μ whose support is contained in K, and such that L is represented as $L(p) = \int p \, d\mu$, for all $p \in [\underline{X}]$.

Partial answers to the $\mathcal{A}$–truncated K–moment problem are known. For example, when K is a closed subset of ^{n} and $\mathcal{A}$ is the set of all monomials up to a certain degree $2d$ or $2d+1$, the existence of such a K–representing measure is proved to be equivalent to the K–positive extendability of L to the set of all polynomials of degree at most $2d+2$, (16, Theorem 2.2). When K is compact and Span($\mathcal{A}$) contains a polynomial that is strictly positive on K, the existence of a K–representing measure is known to be equivalent to the K–positivity of L (see (56, Theorem I, p.129), (16, p. 2710), (19, Theorem 2.2) and (45, Algorithm 4.2)).

Our research deals with an abstract version of the truncated moment problem. We show that all the abovementioned solutions can be considered as particular cases of a general result about the existence of positive extensions of linear functionals to larger linear subspaces containing an element that dominates all the members of the original domain. The scope of our study is also much broader as we consider unital commutative –algebras instead of $[\underline{X}]$ and arbitrary linear subspaces of the algebra in lieu of finite dimensional ones. Thus, our setting is general enough to encompass also infinite dimensional instances of the moment problem, e.g. when the algebra is not finitely generated or when the representing measure is supported on an infinite dimensional vector space.

Let A be a unital commutative -algebra, K a closed subset of the character space of A, and B a linear subspace of A. For a linear functional $L : B \to$, we investigate conditions under which L admits an integral representation with respect to a positive Radon measure supported in K. When A is equipped with a submultiplicative seminorm, we employ techniques from the theory of positive extensions of linear functionals to prove a criterion for the existence of such an integral representation for L. When no topology is prescribed on A, we identify suitable assumptions on A, K, B and L which allow us to construct a seminormed structure on A, so as to exploit our previous result to get an integral representation for L. Our main theorems allow us to extend some well-known results on the Classical Truncated Moment Problem, the Truncated Moment Problem for point processes, and the Subnormal Completion Problem for 2–variable weighted shifts. We also analyze the relation between the Full and the Truncated Moment Problem in our general setting; we obtain a suitable generalization of Stochel's Theorem which readily applies to Full Moment Problems for localized algebras.

Infinite dimensional moment problems have been studied already in the sixties (see e.g. (8; 46; 20; 42; 3; 5; 39; 4; 49)) motivated by fundamental questions in applied areas such as statistical physics and quantum mechanics. Since then there has been an extensive production on the infinite dimensional moment problems appearing in the analysis of interacting particle systems as well as in stochastic geometry, spatial ecology, neural spike trains, heterogeneous materials and random packing (see, e.g., (1; 40; 33; 44; 6; 53; 55)). Despite the vast literature devoted to the theory of the infinite dimensional moment problem, and more generally of the moment problem on unital commutative algebras (see (32; 21; 2; 7; 22; 41; 25; 51; 27; 28), just to mention a few recent developments), several questions remain open (cf. (26)).

We also consider several applications of our main results, ranging from the Classical TMP to the TMP for point processes, and to the SCP for 2–variable weighted shifts. In particular, in the case of bivariate polynomials, our novel approach allows us to deal with not only the cases of the Rectangular, Triangular, and Sparse Connected TMP, but also with a new hybrid case which includes the presence of infinitely many moments in one of the variables.

REFERENCES

[1] S. Albeverio and F. Herzberg, The moment problem on the Wiener space, *Bull. Sci. Math.* 132(2008), no.1, 7–18.

[2] D. Alpay, P.E.T. Jorgensen and D.P. Kimsey, Moment problems in an infinite number of variables, *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* 18(2015), no. 4, 14 pp.

[3] Yu. M. Berezansky and S.N. Šifrin, A generalized symmetric power moment problem (Russian), *Ukrain. Mat. Ž.* 23(1971), 291–306.

[4] Y.M. Berezansky and Y. G. Kondratiev, *Spectral methods in infinite-dimensional analysis*, vol. II, Naukova Dumka, Kiev, 1988 (in Russian); English translation: Kluwer Academic Publishers, Dordrecht, 1995.

[5] H.J. Borchers and J. Yngvason, Integral representations for Schwinger functionals and the moment problem over nuclear spaces, *Comm. Math. Phys.* 43(1975), no. 3, 255–271.

[6] E.N. Brown, R. Kass, P.P. Mitra, Multiple neural spike train data analysis: state-of-the- art and future challenges, *Nature Neuroscience* 7(2004), 456–471.

[7] E. Caglioti, M. Infusino and T. Kuna, Translation invariant realizability problem on the d-dimensional lattice: an explicit construction, *Electron. Commun. Probab.* 21(2016), no. 45, 9 pp.

[8] A.J. Coleman, Structure of Fermion density matrices, *Rev. Mod. Phys.* 35(1963), 668–686.

[9] R.E. Curto and L.A. Fialkow, Recursiveness, positivity, and truncated moment problems, *Houston J. Math.* 17(1991), 603–635.

[10] R.E. Curto and L.A. Fialkow, Recursively generated weighted shifts and the subnormal completion problem, *Integral Equations Operator Theory* 17(1993), 202–246.

[11] R.E. Curto and L.A. Fialkow, *Solution of the truncated complex moment problem with flat data*, *Memoirs Amer. Math. Soc.*, no. 568, Amer. Math. Soc., Providence, 1996.

[12] R.E. Curto and L.A. Fialkow, *Flat extensions of positive moment matrices: Recursively generated relations*, *Memoirs Amer. Math. Soc*, no. 648, Amer. Math. Soc., Providence, 1998.

[13] R.E. Curto and L.A. Fialkow, The truncated complex K–moment problem, *Trans. Amer. Math. Soc.* 352(2000), 2825–2855.

[14] R.E. Curto and L.A. Fialkow, A duality proof of Tchakaloff's theorem, *J. Math. Anal. Appl.* 269(2002), 519–532.

[15] R.E. Curto and L.A. Fialkow, Truncated K–moment problems in several variables, *J. Operator Theory* 54(2005), 189–226.

[16] R.E. Curto and L.A. Fialkow, An Analogue of the Riesz-Haviland Theorem for the truncated moment problem, *J. Funct. Anal.* 255(2008), 2709–2731.

[17] P.J. di Dio, The multidimensional truncated moment problem: Gaussian and log-normal mixtures, their Carathéodory numbers, and set of atoms, *Proc. Amer. Math. Soc.* 147(2019), 3021–3038.

[18] P. J. di Dio and K. Schmüdgen, The multidimensional truncated Moment Problem: Carathéodory Numbers, *J. Math. Anal. Appl.* 461(2018), 1606–1638.

[19] L.A. Fialkow and J. Nie, The truncated moment problem via homogenization and flat extensions, *J. Funct. Anal.* 263(2012), 1682–1700.

[20] C. Garrod and J.K. Percus, Reduction of the N-particle variational problem, *J. Math. Phys.* 5(1964), 1756–1776.

[21] M. Ghasemi, S. Kuhlmann and M. Marshall, Application of Jacobi's representation theorem to locally multiplicatively convex topological R-algebras, *J. Funct. Anal.* 266(2014), no. 2, 1041–1049.

[22] M. Ghasemi, S. Kuhlmann and M. Marshall, Moment problem in infinitely many variables, *Israel J. Math.* 212(2016), no. 2, 989–1012.

[23] W. Helton and J. Nie, A semidefinite approach for truncated K-moment problem, *Found. Comp. Math.* 12(2012), no. 6, 851–881.

[24] D. Henrion and J.B. Lasserre, GloptiPoly: Global optimization over polynomials with Mathlab and SeDuMi, ACM Trans. Math. Soft. 29(2003), 165–194.

[25] M. Infusino, S. Kuhlmann and M. Marshall, On the determinacy of the moment problem for symmetric algebras of a locally convex space, in *Operator theory in different settings and related applications*, *Oper. Theory Adv. Appl.* 262(2018), 243–250.

[26] M. Infusino and S. Kuhlmann, Infinite dimensional moment problem: open questions and applications, in *Ordered Algebraic Structures and Related Topics*, *Contemporary Mathematics* 697(2017), 187–201.

[27] M. Infusino and T. Kuna, The full moment problem on subsets of probabilities and point configurations, *J. Math. Anal. Appl.* 483(2020), no. 1, 123551.

[28] M. Infusino, S. Kuhlmann, T. Kuna and P. Michalski, Projective limits techniques for the infinite dimensional moment problem, arxiv: 1906.01691.

[29] M. Infusino, T. Kuna, J.L. Lebowitz and E.R. Speer, The truncated moment problem on $\mathbf{N}_0$. *J. Math. Anal. Appl.* 452(2017), 443–468.

[30] D. Kimsey, The subnormal completion problem in several variables, *J. Math. Anal. Appl.* 434(2016), 1504–1532.

[31] D. Kimsey and H. Woerdeman, The truncated matrix-valued K–moment problem on ^{d}, ^{d}, and $\mathbf{T}^d$, *Trans. Amer. Math. Soc.* 365(2013), 5393–5430.

[32] T. Kuna, J. L. Lebowitz and E.R. Speer, Necessary and sufficient conditions for realizability of point processes. *Ann. Appl. Probab.* 21(2011), 1253–1281.

[33] R. Lachieze-Rey and I. Molchanov, Regularity conditions in the realisability problem in applications to point processes and random closed sets, *Ann. Appl. Probab.* 25(2015), no. 1, 116–149.

[34] J.B. Lasserre, Global optimization with polynomials and the problem of moments, *SIAM J. Optim.* 11(2001) 796–817.

[35] J.B. Lasserre, *Moments, Positive Polynomials and Their Applications*, Imperial College Press Optimization Series, vol. 1, Imperial College Press, London, 2010.

[36] M. Laurent, Revisiting two theorems of Curto and Fialkow on moment matrices, Proc. Amer. Math. Soc. 133(2005), 2965–2976.

[37] M. Laurent, Semidefinite representations for finite varieties, Math. Program. 109(2007), Ser. A, 1–26.

[38] M. Laurent and B. Mourrain, A generalized flat extension theorem for moment matrices, *Arch. Math. (Basel)* 93(2009), 87–98.

[39] A. Lenard, States of classical statistical mechanical systems of infinitely many particles. I, *Arch. Rational Mech. Anal.* 59(1975): 219–239.

[40] Y. G. Kondratiev, T. Kuna and M.J. Oliveira, Holomorphic Bogoliubov functionals for interacting particle systems in continuum, *J. Funct. Anal.* 238(2006), no.2: 375–404.

[41] Y. G. Kondratiev, T. Kuna and E. Lytvynov, A moment problem for random discrete measures, *Stochastic Process. Appl.* 125(2015), no. 9, 3541–3569.

[42] H. Kummer, n-Representability Problem for Reduced Density Matrices, *J. Math. Phys.* 8(1967), no. 10, 2063–2081.

[43] M. Marshall, *Positive Polynomials and Sums of Squares*, Math. Surveys & Monographs 146, Amer. Math. Soc., 2008.

[44] D. J Murrell, U. Dieckmann and R. Law, On moment closure for population dynamics in continuous space, *J. Theo. Bio.* 229(2004), 421-–432.

[45] J. Nie, The $\mathcal{A}$–truncated K–moment problem, *Foundations of Computational Mathematics* 14(2014), 1243–1276.

[46] J.K. Percus, The pair distribution function in classical statistical mechanics, in *The Equilibrium Theory of Classical Fluids*, ed. H.L. Frisch and J.L. Lebowitz, Benjamin, New York, 1964.

[47] M. Putinar, The L problem of moments in two dimensions, *J. Funct. Anal.* 94(1990), 288–307.

[48] M. Putinar and F.H. Vasilescu, Solving moment problems by dimensional extension, *Ann. Math.* 149(1999), 1087–1107.

[49] K. Schmüdgen, *Unbounded operator algebras and representation theory*, Oper. Theory: Adv. Appl. 37, Birkhäuser Verlag, Basel 1990.

[50] K. Schmüdgen, *The Moment Problem*, Graduate Texts in Mathematics, Springer, 2017.

[51] K. Schmüdgen, On the infinite dimensional moment problem, *Ark. Mat.* 56(2018), no. 2, 441-–459.

[52] J. Stampfli, Which weighted shifts are subnormal?, *Pacific J. Math.* 17(1966), 367-379.

[53] F.H. Stillinger and S. Torquato, Pair correlation function realizability: Lattice model implications, *J. Phys. Chem. B*, 108(2004), 19589–19594.

[54] J. Stochel, Solving the truncated moment problem solves the moment problem, *Glasgow Math. J.* 43(2001), 335–341.

[55] S. Torquato and F.H. Stillinger, New conjectural lower bounds on the optimal density of sphere packings, *Exp. Math.* 15(2006), no.3, 307–331.

[56] V. Tchakaloff, Formules de cubatures mécanique à coefficients non négatifs, *Bull. Sci. Math.* 81(1957), 123–134.

[57] A. Zalar, The truncated Hamburger moment problem with gaps in the index set, *Intergral Equations Operator Theory* 93(2021), no. 3, article number 22; 36 pp.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On the exact neural approximations of MPC laws ⋆

Filippo Fabiani * **Paul J. Goulart** *

* *Department of Engineering Science, University of Oxford, OX1 3PJ, United Kingdom (e-mail: {filippo.fabiani, paul.goulart}@eng.ox.ac.uk)*

Abstract: In this extended abstract we present a general procedure to quantify the performance of rectified linear unit (ReLU) neural network (NN) controllers that preserve the desirable properties of a designed model predictive control (MPC) scheme. First, by quantifying the approximation error between NN and MPC state-to-input mappings, we establish suitable conditions involving the worst-case error and the Lipschitz constant that guarantee the stability of the closed-loop system. Then, we develop an offline, mixed-integer (MI) optimization-based method to compute those quantities exactly, thus providing an analytical tool to certify the stability and performance of a ReLU-based approximation of an MPC control law.

Keywords: Model predictive control, Neural networks, Linear systems in control theory, Mixed-integer programming.

1. INTRODUCTION

Model predictive control (MPC) is one of the most popular control strategies for linear systems with operational and physical constraints (Rawlings et al., 2017) and is based on the repeated solution of constrained optimal control problems. Although the theory underlying MPC provides practical stability and performance guarantees, implementations of both *implicit* and *explicit* MPC suffer from well-known practical difficulties (i.e., solving optimization problems in real-time and complexity of the resulting piecewise-affine (PWA) controller, respectively).

As a result, the idea of *approximating* an MPC policy using (deep) NNs (Goodfellow et al., 2016) is particularly attractive in view of their universal approximation capabilities (Hornik, 1991). Starting from the pioneering work in (Parisini and Zoppoli, 1995), indeed, interest in NN-based approximations of MPC laws has increased rapidly in recent years (Chen et al., 2018; Hertneck et al., 2018; Karg and Lucia, 2020a,b; Zhang et al., 2021; Maddalena et al., 2020; Paulson and Mesbah, 2020). Despite their computationally demanding offline training requirements, the online evaluation of NN-based approximations to MPC laws is computationally inexpensive, since it only requires the evaluation of an input-output mapping. However, unless one assumes a certain structure, NNs are generally hard to analyze due to their nonlinear, large-scale structure.

In this extended abstract we provide a means to assess the training quality of a ReLU network in replicating the action of an MPC policy. First, by considering the approximation error between a ReLU network and an MPC law, we give sufficient conditions involving the maximal approximation error and the associated Lipschitz constant that guarantee the closed-loop stability of a discrete-time linear time-invariant (LTI) system when the ReLU approximation replaces the MPC law. Successively, we formulate a mixed-integer linear program (MILP) to compute the Lipschitz constant of an MPC policy exactly. Finally, we develop an optimization-based technique to exactly compute the worst-case approximation error and the Lipschitz constant characterizing the approximation error. The outcome is a set of conditions involving the optimal value of two MILPs that are sufficient to allow us to certify the reliability of the ReLU-based controller, thus proposing a unifying theoretical framework for analyzing NN-based approximations of MPC policies (the proofs of the technical results are in (Fabiani and Goulart, 2021)).

⋆ This work was partially supported through the Government's modern industrial strategy by Innovate UK, part of UK Research and Innovation, under Project LEO (Ref. 104781).

2. MOTIVATIONS, PROBLEM FORMULATION AND PRELIMINARY RESULTS

We will consider the problem of stabilizing the constrained, discrete-time, LTI system

$$x^+ = Ax + Bu, \tag{1}$$

with state variable $x \in \mathcal{X}$, control input $u \in \mathcal{U}$, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. We will assume that the constraint sets $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{U} \subseteq \mathbb{R}^m$ are bounded polyhedral. A popular control choice for constrained systems is MPC that requires one to solve, at every iteration, the following multi-parametric quadratic program (mp-QP) over a time horizon of length $T \geq 1$, $\mathcal{T} \coloneqq \{0, \dots, T-1\}$,

$$V_T(x) = \begin{cases} \min\limits_{(v_i)_{i\in\mathcal{T}}} & \frac{1}{2}\|x_T\|_P^2 + \sum\limits_{i\in\mathcal{T}} \frac{1}{2}(\|x_i\|_Q^2 + \|v_i\|_R^2) \\ \text{s.t.} & x_{i+1} = Ax_i + Bv_i,\ i \in \mathcal{T}, \\ & x_i \in \mathcal{X},\ v_i \in \mathcal{U},\ i \in \mathcal{T}, \\ & x_0 = x. \end{cases} \tag{2}$$

Starting from some $x(0) \in \mathcal{X}$, the receding horizon implementation of an MPC law computes an optimal solution $(v_i^\star)_{i\in\mathcal{T}}$, and then applies the control input $u(0) = v_0^\star$ taken from the first part of the optimal sequence (*implicit*

version). This process is then repeated at every time k with initial condition $x_0 = x(k)$, so that the procedure amounts to the implicit computation of a mapping $x \mapsto v_0^\star$, defined as

$$u_{\text{MPC}}(x) \coloneqq v_0^\star(x).$$

Under standard assumptions, it is well known that the associated MPC control law $u(k) = u_{\text{MPC}}(x(k))$ stabilizes the constrained LTI system (1) about the origin (Borrelli et al., 2017; Rawlings et al., 2017) while, at the same time, respecting state and input constraints. In some applications the dynamics of the underlying system may be too fast relative to the time required to compute the solution to the mp-QP in (2). One may then rely on the *explicit* version of the MPC law in (2), i.e., explicit model predictive control (eMPC) (Bemporad et al., 2002), whose closed form expression can be computed offline. The optimal solution mapping $u_{\text{MPC}}(\cdot)$ enjoys a PWA structure (Rockafellar and Wets, 2009, Def. 2.47)) that maps any $x \in \mathcal{X}$ into an affine control action according to some polyhedral partition of $\mathcal{X}$. The partition and associated affine functions for $u_{\text{MPC}}(\cdot)$ can be computed offline, e.g. using MPC Toolbox (Bemporad et al., 2021). However, the computational effort required for this offline computation may itself be too demanding, since the number of regions in the optimal partition can grow exponentially with the number of states and constraints in (2) (Alessio and Bemporad, 2009). In addition, even if computable offline, the online implementation of the explicit solution may require excessive memory storage or processing power.

2.1 Approximation of MPC laws via ReLU networks

The aforementioned limitations motivate the design of an approximation for $u_{\text{MPC}}(\cdot)$ that can be implemented with minimal computation and storage requirements while still maintaining stability and good performance of the closed-loop system. We focus on controllers implemented using ReLU neural networks, which provide a natural means for approximating $u_{\text{MPC}}(\cdot)$ since the output mapping of such a network has a PWA structure (Montufar et al., 2014). An L-layered, feedforward ReLU network that defines a mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ can be formally described by the following recursive equations across layers (Hagan et al., 1997):

$$\begin{cases} x^0 = x, \\ x^{j+1} = \max(W^j x^j + b^j, 0), \ j \in \{0, \dots, L-1\}, \\ F(x) = W^L x^L + b^L, \end{cases} \tag{3}$$

where $x^0 = x \in \mathbb{R}^{n_0}$, $n_0 = n$, is the input to the network, and $W^j \in \mathbb{R}^{n_{j+1} \times n_j}$, $b^j \in \mathbb{R}^{n_{j+1}}$ are the weight matrix and bias vector of the $(j+1)$-th layer, respectively (defined during some offline training phase). The total number of neurons is thus $N \coloneqq \sum_{j=1}^{L} n_j + m$, since $n_{L+1} = m$.

Thus, after training a ReLU network to produce a mapping $u_{\text{NN}} : \mathcal{X} \to \mathbb{R}^m$, $u_{\text{NN}}(x) \coloneqq F(x)$, to approximate $u_{\text{MPC}}(\cdot)$, we ask whether the training was sufficient to ensure stability of the closed-loop system in (1) with piecewise-affine neural network (PWA-NN) controller $u_{\text{NN}}(\cdot)$ in place of $u_{\text{MPC}}(\cdot)$. We characterize next the features of the approximation error function $e(x) \coloneqq u_{\text{NN}}(x) - u_{\text{MPC}}(x)$ that are key to guarantee the stability of the closed-loop system in (1) with the PWA-NN controller $u_{\text{NN}}(\cdot)$.

2.2 Closed-loop stability with neural network controllers

We hence investigate the robust stability of (1) with MPC policy, $u_{\text{MPC}}(\cdot)$, subject to an additive disturbance:

$$x^+ = Ax + Bu_{\text{NN}}(x) = Ax + Bu_{\text{MPC}}(x) + Be(x). \tag{4}$$

For some $\alpha \in \{1, \infty\}$, we assume that there exist finite constants $\bar{e}_\alpha$, $\mathcal{L}_\alpha(e, \mathcal{X}_\infty) \geq 0$ such that $\|e(x)\|_\alpha \leq \bar{e}_\alpha$, for all $x \in \mathcal{X}$, and $\|e(x) - e(y)\|_\alpha \leq \mathcal{L}_\alpha(e, \mathcal{X}_\infty)\|x - y\|_\alpha$, for all x, $y \in \mathcal{X}_\infty$, where the set $\mathcal{X}_\infty$ will be defined later. In §4, we will show how these conditions can be made to hold, providing a MI optimization-based method to compute $\bar{e}_\alpha$ and $\mathcal{L}_\alpha(e, \mathcal{X}_\infty)$ exactly. For the remainder we make the following mild assumption:

Standing Assumption 1. For the LTI system (1) under the action of the MPC based controller u_{MPC}:

i) the origin is exponentially stable;
ii) the mp-QP in (2) is recursively feasible starting from any $x \in \mathcal{X}$. □

By exploiting the optimal cost $V_T(\cdot)$ of the mp-QP in (2), with $\Omega_c \coloneqq \{x \in \mathcal{X} \mid V_T(x) \leq c\}$ denoting the sublevel set $c \coloneqq \max\{a \geq 0 \mid \Omega_a \subseteq \mathcal{X}\}$, we first establish that the closed-loop system in (1) with an approximated MPC law is input-to-state stable (ISS) (Jiang and Wang, 2001) when the worst-case approximation error $\bar{e}_\alpha$ is sufficiently small.

Lemma 1. There exists $\zeta > 0$ such that, if $\bar{e}_\alpha < \zeta$, the system in (1) with $u = u_{\text{NN}}(x)$ converges exponentially to some set $\Omega_b \subset \Omega_c$, for all $x(0) \in \Omega_c$. □

Thus, if the maximal approximation error over $\mathcal{X}$ is strictly smaller than ζ, which is tunable through a nonnegative parameter also affecting the size of Ω_b, then the closed-loop in (1) with PWA-NN controller $u_{\text{NN}}(\cdot)$ is ISS and its state trajectories satisfy the constraints, since Ω_c is robust positively invariant. Now, define $\mathcal{X}_\infty$ as the set of states for which the stabilizing unconstrained linear gain (typically the linear quadratic regulator (LQR)), $\bar{K}_{\text{MPC}} \in \mathbb{R}^{m \times n}$, satisfies both state and control constraints, i.e., $\mathcal{X}_\infty \coloneqq \{x \in \mathcal{X} \mid x(0) = x \in \mathcal{X}, Ax(k) + B\bar{K}_{\text{MPC}}x(k) \in \mathcal{X}, \bar{K}_{\text{MPC}}x(k) \in \mathcal{U}, k \in \mathcal{T}, x(T) \in \mathcal{X}\}$ (maximal output admissible set – see (Gilbert and Tan, 1991)). Within $\mathcal{X}_\infty$, the system (4) still enjoys exponential convergence if the local Lipschitz constant of $e(\cdot)$ meets a certain condition:

Lemma 2. There exists $\vartheta > 0$ such that, if $\mathcal{L}_\alpha(e, \mathcal{X}_\infty) < \vartheta$, the system in (1) with $u = u_{\text{NN}}(x)$ converges exponentially to the origin for all $x(0) \in \mathcal{X}_\infty$. □

Putting the previous results together gives us our main stability result, upon which subsequent requirements on the fidelity of our ReLU-based controllers will be based:

Theorem 2.1. If $\bar{e}_\alpha < \zeta$, $\mathcal{L}_\alpha(e, \mathcal{X}_\infty) < \vartheta$, and $b \geq 0$ is chosen so that $\Omega_b \subseteq \mathcal{X}_\infty$, the system in (1) with $u = u_{\text{NN}}(x)$ converges exponentially to the origin, for all $x(0) \in \Omega_c$. □

In the remainder, we provide a MI optimization-based method to compute the maximum approximation error and the (local) Lipschitz constant of $e(\cdot)$ exactly, thus providing conditions sufficient to certify the stability and performance of a ReLU-based approximation of an MPC control law.

3. EXACTL LIPSCHITZ CONSTANT COMPUTATION VIA MIXED-INTEGER LINEAR PROGRAM

We next develop a method of computing the maximum gain (Darup et al., 2017) (and hence the Lipschitz constant, according to (Gorokhovik et al., 1994, Prop. 3.4)) of the MPC policy $u_{\text{MPC}}(\cdot)$ directly via MI programming.

Note that the maximum gain can also be computed by means of available tools that compute the complete explicit solution to mp-QP in (2) directly, e.g., the MPC Toolbox (Bemporad et al., 2021). However, from (Jordan and Dimakis, 2020) we know that the Lipschitz constant of a ReLU network, which we will use to approximate the MPC policy in (2), can itself be computed through a MILP. We therefore require a technique compatible with the one proposed in (Jordan and Dimakis, 2020), which will also allow us subsequently to compute key quantities characterizing the approximation error $e(\cdot)$, according to §2. For the results we are about to introduce, we make a further standard assumption characterizing the constraints of (2) to rule out pathological cases when computing the (unique) solution to the mp-QP in (2).

Standing Assumption 2. (Linear independence constraint qualification (Borrelli et al., 2017, Def. 2.1)) For all $x \in \mathcal{X}$, the linear independence constraint qualification (LICQ) is assumed to hold for the mp-QP in (2). □

Proposition 1. Suppose $\mathcal{X} \subseteq \mathbb{R}^n$ is a polytope and $K : \mathcal{X} \to \mathbb{R}^{m\times n}$ an affine function. Then computing $\mathcal{L}_\alpha(K, \mathcal{X}) = \max_{x\in\mathcal{X}} \|K(x)\|_\alpha$ amounts to an MILP. □

Proposition 1 says that the norm of a matrix whose entries are affine in $x \in \mathcal{X}$ can be computed through an MILP. We now state the main result of this section, which says that the maximum matrix norm taken over the entire partition induced by $u_{\text{MPC}}(\cdot)$ can also be computed via an MILP:

Theorem 3.2. Computing $\mathcal{L}_\alpha(u_{\text{MPC}}, \mathcal{X})$ amounts to an MILP. □

In Table 1 we contrast numerically our proposed approach in Theorem 3.2 with the solution obtained via the MPC Toolbox (Bemporad et al., 2021). Here, the column "# of $\mathcal{R}_\mathcal{A}$" reports the number of polyhedral partitions characterizing eMPC.

4. QUANTIFYING THE APPROXIMATION QUALITY OF PIECEWISE-AFFINE NEURAL NETWORKS

We can now develop computational results that ensure the stability of a ReLU-based control policy u_{NN} constructed based on approximation of a stabilizing MPC law u_{MPC}.

Since the MPC policy u_{MPC} is designed to (exponentially) stabilize the LTI system in (1) to the origin, then we may expect that the ReLU based policy should also be stabilizing if the approximation error $e(\cdot) = u_{\text{NN}}(\cdot) - u_{\text{MPC}}(\cdot)$ is sufficiently small. This error function is the difference of PWA functions, and so also PWA (Gorokhovik et al., 1994, Prop. 1.1). Thus, it can similarly be shown to be bounded and Lipschitz continuous on $\mathcal{X}$, and we can therefore apply the results of §2 to find conditions under which stability is preserved. Whether or not the error can be made sufficiently small to do so depends on both the amount (and quality) of training data and the complexity of the network (i.e., number of neurons and layers).

Theorem 4.3. The approximation error $e(\cdot) = u_{\text{NN}}(\cdot) - u_{\text{MPC}}(\cdot)$ has the following properties:

i) The maximal error $\max_{x\in\mathcal{X}}\|e(x)\|_\alpha =: \bar{e}_\alpha$ can be computed by solving an MILP;
ii) The Lipschitz constant $\mathcal{L}_\alpha(e, \mathcal{X})$ can be computed by solving an MILP. □

By combining known results available in the machine learning literature and the ones developed in this extended abstract, Theorem 4.3 provides an offline, optimization-based procedure to compute exactly both the worst-case approximation error between the PWA mappings associated with the ReLU network in (3) and the MPC law in (2), as $\|e(x)\|_\alpha \le \bar{e}_\alpha$, for all $x \in \mathcal{X}$, and the associated Lipschitz constant over $\mathcal{X}$, $\mathcal{L}_\alpha(e, \mathcal{X})$, for $\alpha \in \{1, \infty\}$. These quantities are precisely of the type required to apply the stability results of §2, and thus allow us to certify the reliability, in terms of stability of the closed-loop system, of a ReLU-based approximation of a given MPC law.

We finally stress that, while $u_{\text{NN}}(\cdot)$ can guarantee state constraint satisfaction for any initial state $x(0) \in \Omega_c$ if some prescribed conditions are met (§2), the input constraints may not be satisfied. This issue can be accommodated either during the training phase by adopting the available techniques in, e.g., (Chen et al., 2018; Karg and Lucia, 2020a,b; Paulson and Mesbah, 2020), or directly after the training process through output verification methods (Bunel et al., 2018; Fazlyab et al., 2020).

5. CONCLUSION

We have shown that the design of ReLU-based approximations with provable stability guarantees require one to construct and solve two MILPs offline, whose associated optimal values characterize key quantities of the approximation error. We have provided a systematic way to encode the maximal gain of a given MPC law through binary and continuous variables subject to MI constraints. This optimization-based result is compatible with existing results from the machine learning literature on computing the Lipschitz constant of a trained ReLU network. Taken together they provide sufficient conditions to assess the reliability, in terms of stability of the closed-loop system, of a given ReLU-based approximation of an MPC scheme.

REFERENCES

Alessio, A. and Bemporad, A. (2009). A survey on explicit model predictive control. In *Nonlinear model predictive control*, 345–369. Springer.

Bemporad, A. and Filippi, C. (2003). Suboptimal explicit receding horizon control via approximate multiparametric quadratic programming. *Journal of Optimization Theory and Applications*, 117(1), 9–38.

Bemporad, A., Morari, M., Dua, V., and Pistikopoulos, E.N. (2002). The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1), 3–20.

Bemporad, A., Ricker, N.L., and Morari, M. (2021). Model predictive control toolbox. *User's Guide, Version*, 2.

Borrelli, F., Bemporad, A., and Morari, M. (2017). *Predictive control for linear and hybrid systems*. Cambridge University Press.

Table 1. Numerical results for the computation of $\mathcal{L}_\alpha(u_{\text{MPC}}, \mathcal{X})$, $\alpha \in \{1, \infty\}$

Reference	MPC Toolbox (Bemporad et al., 2021)				MILP Th. 3.2 + Prop. 1			
	# of $\mathcal{R}_\mathcal{A}$	CPU time [s]	$\mathcal{L}_1$	$\mathcal{L}_\infty$	$\mathcal{L}_1$	CPU time [s]	$\mathcal{L}_\infty$	CPU time [s]
(Darup, 2014)	99	1.24	11.67	16.6	11.7	0.51	16.1	0.45
(Gutman and Cwikel, 1987)	111	1.56	7.99	11.98	8.00	0.56	12.00	0.69
(Darup and Cannon, 2016)	27	1.18	0.49	0.5	0.5	2.48	0.5	3.89
(Gutman and Cwikel, 1987)	317	23	1.27	1.88	1.27	10.8	1.88	16.1
(Bemporad and Filippi, 2003)	105	1.34	2.39	3.1	2.39	0.99	3.1	0.98
(Jones and Morari, 2009)	729	117.8	1.52	1.76	1.53	8.4	1.77	7.9
(Markolf and Stursberg, 2021)	15	5.46	1.66	1.66	1.66	0.14	1.66	0.14

Bunel, R.R., Turkaslan, I., Torr, P., Kohli, P., and Mudigonda, P.K. (2018). A unified view of piecewise linear neural network verification. In *Advances in Neural Information Processing Systems*, volume 31, 4795–4804.

Chen, S., Saulnier, K., Atanasov, N., Lee, D.D., Kumar, V., Pappas, G.J., and Morari, M. (2018). Approximating explicit model predictive control using constrained neural networks. In *2018 Annual American control conference (ACC)*, 1520–1527. IEEE.

Darup, M.S. (2014). *Numerical methods for the investigation of stabilizability of constrained systems*. Ph.D. thesis, Ruhr-Universität Bochum.

Darup, M.S. and Cannon, M. (2016). Some observations on the activity of terminal constraints in linear MPC. In *2016 European Control Conference (ECC)*, 770–775. IEEE.

Darup, M.S., Jost, M., Pannocchia, G., and Mönnigmann, M. (2017). On the maximal controller gain in linear MPC. *IFAC-PapersOnLine*, 50(1), 9218–9223.

Fabiani, F. and Goulart, P.J. (2021). Reliably-stabilizing piecewise-affine neural network controllers. *IEEE Transactions on Automatic Control*. (Under review – also available at `https://arxiv.org/abs/2111.07183`).

Fazlyab, M., Morari, M., and Pappas, G.J. (2020). Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. *IEEE Transactions on Automatic Control*. (In press).

Gilbert, E.G. and Tan, K.T. (1991). Linear systems with state and control constraints: The theory and application of maximal output admissible sets. *IEEE Transactions on Automatic control*, 36(9), 1008–1020.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gorokhovik, V.V., Zorko, O.I., and Birkhoff, G. (1994). Piecewise affine functions and polyhedral sets. *Optimization*, 31(3), 209–221. doi:10.1080/02331939408844018.

Gutman, P.O. and Cwikel, M. (1987). An algorithm to find maximal state constraint sets for discrete-time linear dynamical systems with bounded controls and states. *IEEE Transactions on Automatic Control*, 32(3), 251–254.

Hagan, M.T., Demuth, H.B., and Beale, M. (1997). *Neural network design*. PWS Publishing Co.

Hertneck, M., Köhler, J., Trimpe, S., and Allgöwer, F. (2018). Learning an approximate model predictive controller with guarantees. *IEEE Control Systems Letters*, 2(3), 543–548.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2), 251–257.

Jiang, Z.P. and Wang, Y. (2001). Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6), 857–869.

Jones, C.N. and Morari, M. (2009). Approximate explicit MPC using bilevel optimization. In *2009 European control conference (ECC)*, 2396–2401. IEEE.

Jordan, M. and Dimakis, A.G. (2020). Exactly computing the local Lipschitz constant of ReLU networks. In *Advances in Neural Information Processing Systems*, volume 33, 7344–7353.

Karg, B. and Lucia, S. (2020a). Efficient representation and approximation of model predictive control laws via deep learning. *IEEE Transactions on Cybernetics*, 50(9), 3866–3878.

Karg, B. and Lucia, S. (2020b). Stability and feasibility of neural network-based controllers via output range analysis. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 4947–4954. IEEE.

Maddalena, E.T., Moraes, C.G.d.S., Waltrich, G., and Jones, C.N. (2020). A neural network architecture to learn explicit MPC controllers from data. *IFAC-PapersOnLine*, 53(2), 11362–11367.

Markolf, L. and Stursberg, O. (2021). Polytopic input constraints in learning-based optimal control using neural networks. In *2021 20th European Control Conference (ECC)*. IEEE.

Montufar, G.F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, 27, 2924–2932.

Parisini, T. and Zoppoli, R. (1995). A receding-horizon regulator for nonlinear systems and a neural approximation. *Automatica*, 31(10), 1443–1451.

Paulson, J.A. and Mesbah, A. (2020). Approximate closed-loop robust model predictive control with guaranteed stability and constraint satisfaction. *IEEE Control Systems Letters*, 4(3), 719–724.

Rawlings, J.B., Mayne, D.Q., and Diehl, M. (2017). *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing.

Rockafellar, R.T. and Wets, R.J.B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Zhang, X., Bujarbaruah, M., and Borrelli, F. (2021). Near-optimal rapid MPC using neural networks: A primal-dual policy learning framework. *IEEE Transactions on Control Systems Technology*, 29(5), 2102–2114.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Lower bound performance for averaging algorithms in open multi-agent systems ⋆

Charles Monnoyer de Galland * **Julien M. Hendrickx** *

* *ICTEAM Institute, UCLouvain, B-1348 Louvain-la-Neuve, Belgium.*
(e-mail: {charles.monnoyer, julien.hendrickx}@uclouvain.be).

Abstract: We derive fundamental limitations on the performance that can be achieved by intrinsic averaging algorithms in open multi-agent systems, which are systems subject to random arrivals and departures of agents. Each agent holds an intrinsic value, and their objective is to collaboratively estimate the average of the values of the agents presently in the system. We provide a lower bound on the expected Mean Square Error for such algorithms where we assume that the size of the system remains constant. Our derivation is based on the error obtained with an algorithm that achieves optimal performance under a set of restrictions on the way agents obtain information about one another. This error represents a lower bound on the error obtained with any other algorithm that can be implemented under the same restrictions. This approach is then applied to derive lower bounds on the performance of the well-known Gossip algorithm by considering restrictions that allow implementing it.

Keywords: Systems theory; control, Probability theory and stochastic processes, Average consensus, Open systems

1. PRELIMINARY NOTE

This work is a resubmission of the extended abstract that was accepted for MTNS 2020.

2. INTRODUCTION

Multi-agent systems are being largely studied in various application fields, including *e.g.*, formation control or opinion dynamics, especially for their robustness, flexibility and scalability. Yet, most results obtained in that framework build on the assumption that the composition of the systems remains unchanged throughout the whole process. This assumption is getting increasingly challenged by the growing size of the systems, as it slows down the process, and makes small individual probabilities of arrivals or departures within the system non-negligible. Such systems, called *open*, can also naturally arise when the process is slow enough or chaotic by nature and where communications can be difficult or happen at a time-scale similar to that of the process, e.g. collaborative multi-vehicle systems where vehicles share a stretch of road for a time before heading to different destinations.

In such configurations, analyses and algorithm design become challenging as the state, size, and at some extent objective pursued by the system vary over time with arrivals and departures of agents. Those incessant perturbations require designing algorithms able to deal with a variable objective, and prevent them to achieve the usual convergence. On top of that, results obtained for closed systems do not easily extend to open ones, see e.g. Hendrickx and Martin (2016); Abdelrahim et al. (2017). Little work exists around open systems, including simulation-based analyses performed by Sen and Chakrabarti (2013) or through size-independent quantities in Hendrickx and Martin (2016), and algorithm design for MAX consensus from Abdelrahim et al. (2017). In this work, we extend the results presented in Monnoyer de Galland and Hendrickx (2019) by proposing a general formulation of fundamental performance limitations for algorithms that can be implemented under some restrictions on the exchange of information within the system, and apply it to the Gossip algorithm.

⋆ ICTEAM institute, UCLouvain (Belgium). This work was supported by "Communauté française de Belgique - Actions de Recherche Concertées". C. M. is a FRIA fellow (F.R.S.-FNRS). This work is supported by the "*RevealFlight*" ARC at UCLouvain. Email adresses: charles.monnoyer@uclouvain.be, julien.hendrickx@uclouvain.be.

3. PROBLEM STATEMENT

We consider a multi-agent system constituted of N agents labelled from 1 to N, where each agent j owns a constant intrinsic value x_j drawn from some zero-mean distribution of variance σ^2. Every agent j is randomly replaced according to a Poisson process of individual rate $\lambda_r^{(j)}$ resulting in the erasure of the memory of the agent, and to the attribution of a new value x_j to that agent, so that the system is open and its size is constant. Hence the value held by the agents can be seen as a time varying quantity $x_j(t)$ which is modified at replacements. Observe however that our model differs from fixed-size systems with time-varying states and/or topologies: as replacements also induce the erasure of the memory of the replaced agent, they actually correspond to a new agent connecting for the first time to the system, without having access to any information about the past of the process.

Moreover, we consider the agents can collect information about each other at some times in a distributed manner through local interactions. Based on these, their objective is to collaboratively compute the time-varying average of the intrinsic values of the system defined as $\bar{x}(t) := \frac{1}{N}\sum_{j=1}^{N} x_j(t)$. In this work we study the performance of algorithms solving this problem which agents can implement in that setting.

One such algorithm is the well-known Gossip algorithm (see e.g. Boyd et al. (2006)) designed for closed systems. This algorithm relies on pairwise interactions in order to update an estimation of the average: an agent i initializes its estimate to its own value $y_i(t) = x_i$, and then updates it each time it interacts with some agent j at some time t as follows:

$$y_i(t^+) = y_j(t^+) = \frac{y_i(t^-) + y_j(t^-)}{2}. \tag{1}$$

Unlike closed systems, open systems do not allow averaging algorithms to converge to the exact average, and we need to design an alternative way to measure their performance. One possible standard criterion, which we consider in this work, is the Mean Square Error (MSE) defined as

$$C(t) := \frac{1}{N}\sum_{i=1}^{N} \left(\bar{x}(t) - y_i(t)\right)^2. \tag{2}$$

More precisely, we derive lower bounds on

$$\mathbb{E}\left[C(t)\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{x}(t) - y_i(t)\right)^2\right] \tag{3}$$

for a set of algorithms that can be implemented under some restrictions on the way agents obtain information about one another. For that purpose, we will evaluate the performance achieved by an algorithm that is provably optimal in a more favorable setting than what is typically allowed. To that end we provide agents with additional knowledge about the dynamics of the system, such as its size and the way replacements and interactions take place, or the distribution from which their values are drawn. The bounds we obtain are then *fundamental performance limitations* for any algorithms that can be implemented under the same restrictions, even if they do not make use of all the available information, and are thus a quality criterion for those algorithms.

In what follows, we propose a generic expression for the lower bound on (3) provided a set of restrictions on the way agents obtain information about one another. We then define such sets that allow implementing the Gossip algorithm presented in equation (1) and derive the corresponding bounds, which are thus valid lower bounds on the performance of that algorithm.

4. MAIN RESULT

In order to derive a lower bound on (3), we define the algorithm that achieves optimal performance under restrictions on the way agents acquire information about each other. This leads to the following definition of that algorithm:

$$y_i^*(t) := \arg\min_{y_i(t)} \left\{ \mathbb{E}\left[\left(\bar{x}(t) - y_i(t)\right)^2 | \omega_i^*(t)\right] \right\}, \tag{4}$$

where $\omega_i^*(t)$ refers to the set containing all the information potentially accumulated by agent i at time t about the other agents in the system under a set of restrictions noted $*$, which characterizes the way that information is obtained by i.

By computing the MSE of algorithm (4), one can by definition obtain a lower bound on that of any other algorithm that can be implemented under the restrictions $*$. This computation relies on a decomposition of the MSE allowing the reduction of its analysis to that of the error of estimation of a single agent j made by another agent i, which is entirely characterized by the most recent information about that agent j available to i. It ultimately leads to

$$\mathbb{E}\left[C(t)\right] \geq \frac{1}{N^3}\sum_{i=1}^{N}\sum_{j=1}^{N} \left(1 - \int_0^t f_{j\to(i)}^{t,*}(s) e^{-2\lambda_r^{(j)} s}\, \mathrm{d}s\right) \sigma^2, \tag{5}$$

where $f_{j\to(i)}^{t,*}(s)$ is a probability density function which characterizes the age of the most recent information about agent j available to i at time t under the restrictions $*$. Observe that this function encapsulates the way information travels in the system, so that one can virtually model any specific graph topology by properly defining it.

The bound (5) is thus a time-dependent lower bound on the MSE of any algorithm that can be implemented under a given set of restrictions $*$. This bound is generic, and properly defining such restrictions $*$ allows instantiating $f_{j\to(i)}^{t,*}(s)$ and then deriving a proper expression for (5).

One can also refine that expression by considering a more particular class of open systems. Typically, one could consider systems characterized by a common replacement rate $\lambda_r^{(j)} = \lambda_r$ and by a probability density function independent of the agents and the time $f_{j\to(i)}^{t,*}(s) = f^*(s)$. Exploiting these properties and taking the limit as $t \to \infty$, a few algebraic steps starting from (5) yield the following reduced expression in steady state for this class of systems:

$$\liminf_{t\to\infty} \mathbb{E}\left[C(t)\right] \geq \frac{N-1}{N^2}\left(1 - \int_0^\infty f^*(s) e^{-2\lambda_r s}\, \mathrm{d}s\right)\sigma^2. \tag{6}$$

5. APPLICATION TO THE GOSSIP ALGORITHM

We finally propose several restriction sets that allow implementing the Gossip algorithm in order to instantiate (5) and (6). This will lead to properly defined bounds that are valid for the Gossip algorithm, and for any other algorithm that can be implemented under those restrictions. We focus on restrictions allowing the agents to gather information through random pairwise interactions happening according to a Poisson process of pairwise rate λ_c (*i.e.* a given pair of agents interacts on average λ_c times per unit of time). One can show that those restrictions indeed allow implementing the Gossip algorithm.

Ping restrictions: The first model relies on a strong assumption which corresponds to the presence of a central unit perfectly knowing the state of the system at all times, and to the absence of erasure of an agent's memory at its replacements. At a communication between two agents, they share what they know, and learn the exact state of the whole system at that time from the central unit. This model is called "*Ping*" in reference to the software testing the reachability of machines in a network, and the corresponding bound is given in steady state and under assumptions of symmetry by the following expression:

$$\liminf_{t\to\infty} \mathbb{E}\left[C(t)\right] \geq \frac{N-1}{N^2}\left(\frac{1}{1 + \frac{N-1}{2}\frac{\lambda_c}{\lambda_r}}\right)\sigma^2. \tag{7}$$

Infection restrictions: The second model is less permissive as it considers that agents only exchange everything they know at the time they communicate to build their estimates. One observes that the travel of information under those restrictions reduces to an infection process where the disease is the information about an agent, and the infection rate is the communication rate λ_c. Two variants of that model arise: respectively with and

without healing (which corresponds to the memory erasure at the replacement of an agent) that respectively lead to the "*SIS*" and "*SI*" models. One then obtains the steady state bound (8) under symmetry assumptions, where w and A contain the information related to the continuous time Markov chains defining the infection process, either for the SI or SIS model:

$$\liminf_{t\to\infty} \mathbb{E}\left[C(t)\right] \geq \frac{N-1}{N^2}\left(1 - w^T A\left(2\lambda_r - A\right)^{-1}\mathbf{e_1}\right)\sigma^2. \quad (8)$$

Those bounds are depicted in Fig. 1, which compares the performance of the Gossip algorithm simulated for ten agents with all the bounds that were obtained above in steady state.

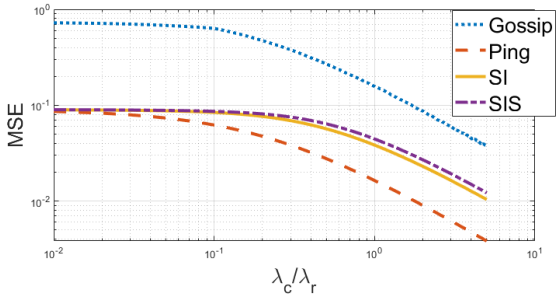


Fig. 1. Steady state MSE comparison showing the validity of (6) for the Gossip performance (blue), where *Ping*, *SI* and *SIS* refer to valid models to instantiate $f^*(s)$ in (6).

It appears that tighter restrictions lead to higher lower bounds on the Mean square Error. More precisely, working with restrictions closer to what the Gossip algorithm actually allows provides a bound that is closer to the actual performance of that algorithm. Moreover, one observes that even though there is a gap between the performance of the Gossip algorithm (in blue) and the different bounds, the general behavior of the MSE is qualitatively well captured. The Gossip is indeed one of the most naive algorithms that can be implemented for achieving average consensus, and does not take into account the openness of the system nor some provided information such as identifiers in its implementation. Hence, it appears that most of the information that was provided to the agents while computing the bounds may be unnecessary in the design of efficient averaging algorithms in open systems in the sense of the Mean Square Error.

6. CONCLUSION

We considered the possibility for agents to join and leave the system in the study of multi-agent systems, and highlighted several challenges that arise in that framework. In particular, this property prevents algorithms to achieve the usual convergence, making their design and analysis challenging.

We focused on averaging algorithms in open systems, for which we obtained a generic formulation of fundamental limitations on their performance given restrictions on the way information is exchanged within the system. Properly defining these restrictions allows deriving lower bounds on the performance of algorithms that are implemented under those restrictions, and thus serve as a quality criterion for them. This was then performed for restrictions allowing implementing the Gossip algorithm, leading to performance limitations for that algorithm.

It appears that defining more restrictive constraints on the information exchange leads to tighter bounds when evaluating the performance of an algorithm. Interestingly, even naive algorithms such as the Gossip show satisfying performance in comparison with rather strong bounds, questioning the possible impact of several parameters that were assumed to be known by the agents when building the bounds, such as identifiers.

Finally, the generalization of our results to systems of time-varying size $N(t)$, *i.e.*, subject to decoupled arrivals and departures is a very interesting yet challenging follow-up for this work. Most on the present analysis relies on the assumption that agents know the system size N, which would be a very strong assumption if that size changes as it gives a significant amount of information about the system. More generally, this extension would require estimating that size in parallel, which is a significant challenge left for future research.

REFERENCES

Abdelrahim, M., Hendrickx, J.M., and Heemels, W.M. (2017). Max-consensus in open multi-agent systems with gossip interactions. In *Proceedings of the 56th IEEE CDC*. doi: 10.1109/CDC.2017.8264362.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms. *IEEE/ACM Trans. Netw.*, 14(SI), 2508–2530. doi:10.1109/TIT.2006.874516. URL `https://doi.org/10.1109/TIT.2006.874516`.

Hendrickx, J.M. and Martin, S. (2016). Open multi-agent systems: Gossiping with deterministic arrivals and departures. In *Proceedings of the 56th IEEE CDC*, 1094–1101. doi: 10.1109/ALLERTON.2016.7852357.

Monnoyer de Galland, C. and Hendrickx, J.M. (2019). Lower bound performances for average consensus in open multi-agent systems (extended version). *arXiv e-prints*, arXiv:1909.02475.

Sen, P. and Chakrabarti, B.K. (2013). *Sociophysics: an introduction*. Oxford University Press.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Characterization and computation of the strong $\mathcal{H}_2$ norm for delay differential algebraic systems ⋆

Wim Michiels * **Marco A. Gomez** ** **Sébestien Mattenet** ***
Vittorio De Iuliis **** **Raphael Jungers** ***

* *Department of Computer Science, KU Leuven, 3001 Heverlee, Belgium (e-mail: Wim.Michiels@cs.kuleuven.be),*
** *Department of Mechanical Engineering, DICIS, Universidad de Guanajuato, 36885 Salamanca, Gto., México, (e-mail: m.galvarez@outlook.com),*
*** *ICTEAM Institute, Université Catholique de Louvain, 3081 Louvain-la-Neuve, Belgium (e-mail: raphael.jungers@uclouvain.be).*
**** *Department of Information Engineering, Computer Science and Mathematics, Università degli Studi dell'Aquila, L'Aquila, Italy (e-mail: vittorio.deiuliis@univaq.it).*

Abstract: The $\mathcal{H}_2$ norm of an exponentially stable system described by Delay Differential Algebraic Equations (DDAEs) might be infinite, due to the existence of hidden feedthrough terms and it might become infinite as a result of infinitesimal changes to the delay parameters. We first introduce the notion of strong $\mathcal{H}_2$ norm of semi-explicit DDAEs, a robustified measure that takes into account delay perturbations, and we analyze its properties. Next, we discuss necessary and sufficient finiteness criteria for the strong $\mathcal{H}_2$ norm in terms of a frequency sweeping test over a hypercube, and in terms of a finite number of equalities involving multi-dimensional powers of a finite set of matrices. Finally, we show that if the $\mathcal{H}_2$ norm of the DDAE is finite, it is possible to construct an exponentially stable neutral delay-differential equation which has the same transfer matrix as the DDAE, without any need for differentiation of inputs or outputs. This connected with a neutral system enables the framework of Lyapunov matrices for computing the $\mathcal{H}_2$ norm.

1. NOTION OF STRONG $\mathcal{H}_2$ NORM

We consider systems described by semi-explicit linear delay-differential algebraic equations (also called coupled delay differential -difference equations) of the form

$$\begin{aligned}\frac{d}{dt}x_1(t) =& A_0^{(11)}x_1(t) + A_0^{(12)}x_2(t) + \sum_{j=1}^{m} A_j^{(11)}x_1(t-h_i)\\ &+ \sum_{j=1}^{m} A_j^{(12)}x_2(t-h_j) + B_1u(t)\\ x_2(t) =& A_0^{(21)}x_1(t) + \sum_{j=1}^{m} A_j^{(21)}x_1(t-h_j)\\ &+ \sum_{j=1}^{m} A_j x_2(t-h_j) + Bu(t)\\ y(t) =& C_1x_1(t) + Cx_2(t),\end{aligned} \tag{1}$$

where $x_1(t) \in \mathbb{R}^r$, $x_2(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^{n_i}$ and $y(t) \in \mathbb{R}^{n_o}$ are state-variables, inputs and outputs at time t. The delays are denoted by $\boldsymbol{h} := (h_1, \ldots, h_m)$.

⋆ This work was supported by project C14/17/072 the KU Leuven Research Council and project G092721N of the Research Foundation-Flanders (FWO - Vlaanderen).

The $\mathcal{H}_2$ norm is an important performance measure in the field of control theory. For an exponentially stable system of the form (1), it is defined as

$$\|G\|_{\mathcal{H}_2} := \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{tr}(G^*(\imath\omega)G(\imath\omega))d\omega},$$

where G is the transfer matrix of system (1).

In contrast to other classes of systems, the $\mathcal{H}_2$ norm of system (1) might be infinite even if the system is exponentially stable, as the DDAE formulation might hide a nontrivial feedthrough term from $u(t)$ to $y(t)$. Furthermore, the function

$$\mathbb{R}^m_{\geq 0} \ni \boldsymbol{h} \mapsto \|G(\cdot;\boldsymbol{h})\|_{\mathcal{H}_2}$$

may not be continuous, even if the system is strongly exponentially stable, as illustrated with the following example.

Example 1. Consider a system of the form

$$\begin{aligned}x_1(t) &= -x_1(t) + u(t),\\ x_2(t) &= A_1x_2(t-h_1) + A_2x_2(t-h_2) + A_3x_2(t-h_3)\\ &\quad + Bu(t),\\ y(t) &= x_1(t) + Cx_2(t),\end{aligned}$$

with matrices

$$A_1 = \begin{pmatrix} 0 & 0 & 0\\ \frac{1}{4} & 0 & -\frac{1}{4}\\ 0 & 0 & 0\end{pmatrix},\ A_2 = \begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & \frac{1}{8}\end{pmatrix},$$

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$$A_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{32} \\ 0 & 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, C = (0\ 1\ 0)\,.$$

The characteristic equation is given by

$$(s+1)\left(1 - \frac{1}{8}e^{-sh_2}\right) = 0,$$

from which we conclude exponential stability for all delay values. The transfer function of the system is given by

$$G(s) = \frac{1}{s+1} - \frac{1}{4}\frac{\left(e^{-s(h_1+h_2)} - e^{-sh_3}\right)}{8 - e^{-sh_2}}.$$

It is clear from this expression that $\|G\|_{\mathcal{H}_2}$ is finite if and only if $h_3 = h_1 + h_2$, while otherwise we have $\|G\|_{\mathcal{H}_2} = +\infty$. Thus, $\|G\|_{\mathcal{H}_2}$ has a discontinuity at each tuple (h_1, h_2, h_3) for which $h_3 = h_1 + h_2$. ⋄

When defining

$$\mathcal{B}(\boldsymbol{h}, \varepsilon) := \{\boldsymbol{\vartheta} \in \mathbb{R}^m_{\geq 0} : \|\boldsymbol{\vartheta} - \boldsymbol{h}\| < \varepsilon\},$$

the strong $\mathcal{H}_2$ norm of a strongly explonentially stable system of the form 1 is defined as

$$|||G(\cdot;\boldsymbol{h})|||_{\mathcal{H}_2} := \lim_{\varepsilon \to 0+} \sup\{\|G(\cdot;\boldsymbol{h}_\varepsilon)\|_{\mathcal{H}_2} : \boldsymbol{h}_\varepsilon \in \mathcal{B}(\boldsymbol{h}, \varepsilon)\}.$$

In order to characterize its finiteness, we define matrix polynomials $P_{k_1,\ldots,k_m}(A_1, \ldots, A_m)$, with $k_j \in \mathbb{Z}_{\geq 0}, j = 1, \ldots, m$, which are recursively defined through the following expressions:

$$P_{0,\ldots,0}(A_1, \ldots, A_m) := I,$$

$$\begin{aligned} P_{k_1,\ldots,k_m}(A_1, \ldots, A_m) &:= A_1 P_{k_1-1,k_2,\ldots,k_m}(A_1, \ldots, A_m) \\ &+ A_2 P_{k_1,k_2-1,\ldots,k_m}(A_1, \ldots, A_m) + \ldots + \\ &+ A_m P_{k_1,k_2,\ldots,k_m-1}(A_1, \ldots, A_m) \end{aligned}$$

and

$P_{k_1,\ldots,k_m}(A_1, \ldots, A_m) := 0$ if any $k_j \in \mathbb{Z}_{<0}, j = 1, \ldots, m$.

For instance, for $m = 2$ and $k_1 + k_2 \leq 3$, these matrix polynomials are

$$\begin{aligned} &P_{0,0} = I, \\ &P_{1,0} = A_1, \quad P_{0,1} = A_2, \\ &P_{2,0} = A_1^2, \quad P_{1,1} = A_1A_2 + A_2A_1, \quad P_{0,2} = A_2^2, \\ &P_{3,0} = A_1^3, \quad P_{2,1} = A_1^2A_2 + A_1A_2A_1 + A_2A_1^2, \\ &P_{1,2} = A_1A_2^2 + A_2A_1A_2 + A_2^2A_1, \quad P_{0,3} = A_2^3. \end{aligned}$$

We can now formulate the following result.

Theorem 1. Assume that system (1) is strongly stable. The following statements are equivalent.

(1) The strong $\mathcal{H}_2$ norm of (1) is finite.
(2) Condition

$$C\left(I - \sum_{j=1}^m A_j e^{-\imath\vartheta_j}\right)^{-1} B = 0, \tag{2}$$

$\forall(\vartheta_1, \ldots, \vartheta_m) \in [0,\ 2\pi]^m$, is satisfied, with $\imath = \sqrt{-1}$.
(3) Conditions

$$CP_{k_1,\ldots,k_m}(A_1, \ldots, A_m)B = 0, \tag{3}$$

$\forall(k_1, \ldots, k_m) \in \mathbb{Z}^m_{\geq 0}$ are satisfied.
(4) Condition (3) hold $\forall(k_1, \ldots, k_m) \in \mathbb{Z}^m_{\geq 0}$ such that

$$\sum_{j=1}^m k_j < n. \tag{4}$$

Furthermore, if the strong $\mathcal{H}_2$ norm of (1) is finite, it equals its $\mathcal{H}_2$ norm.

Determining whether (2) holds, corresponds to checking a semi-infinite equality over a hyper-cube, that can be reduced to checking an infinite but countable number of equalities (3) and finally to a finite number of equalities determined by (4). As shall be discussed in the presentation, the equivalence between statements (1) and (2) follows from an analysis of the feedthrough terms in the transfer matrix, the equivalence between (1) and (3) from an explicit expression of the impulse response. Finally, the restriction (4) in the numbers of equalities to be checked stems from the observation that $\left\{P_{k_1,\ldots,k_m}(A_1, \ldots, A_m) : \sum_{j=1}^m k_j = \ell\right\}$, with $\ell \geq 0$, are the coefficients of the expansion of the multi-variable matrix polynomial $(A_1z_1 + \ldots + A_mz_m)^\ell$, followed by an application of the Cayley-Hamilton theorem.

It is at this point not yet clear how to compute the $\mathcal{H}_2$ norm whenever it is finite. These observations motivate the developments in the next section.

2. TRANSFORMATION TO A NEUTRAL TYPE SYSTEM

For the sake of a clear presentation, let us first recall the two approaches that exist to perform the transformation of (1) to a delay equation of neutral type, which we will refer to in what follows as *regularization mechanisms* (RMs):

RM1 If $B = 0$, then one can apply the operator $\left(\frac{d}{dt} + I\right)$ to the second set of equations, leading to

$$\begin{aligned} &\dot{x}_2(t) - \sum_{i=1}^m A_i\dot{x}_2(t-h_i) - \sum_{i=0}^m A_i^{(21)}\dot{x}_1(t-h_i) = \\ &- x_2(t) + \sum_{i=1}^m A_i x_2(t-h_i) + \sum_{i=0}^m A_i^{(21)} x_1(t-h_i), \end{aligned}$$

which corresponds to a multiplication with $(s+1)$ in the frequency domain. The reason for combining differentiation with addition to the original equation lies in the preservation of internal stability.

RM2 If $C = 0$, one can define a new variable $\hat{x}_2$ via the stable differential equation $\dot{\hat{x}}_2 + \hat{x}_2 = x_2$ and subsequently substituting x_2 in (1), leading to

$$\begin{aligned} \dot{x}_1(t) - \sum_{i=0}^m A_i^{(12)}\dot{\hat{x}}_2(t-h_i) &= \sum_{i=0}^m A_i^{(11)} x_1(t-h_i) \\ &+ \sum_{i=0}^m A_i^{(12)}\hat{x}_2(t-h_i) + B_1u(t) \\ \dot{\hat{x}}_2(t) - \sum_{i=1}^m A_i\dot{\hat{x}}_2(t-h_i) &= -\hat{x}_2(t) + \sum_{i=1}^m A_i\hat{x}_2(t-h_i) \\ &+ \sum_{i=0}^m A_i^{(21)} x_1(t-h_i) + Bu(t), \\ y(t) &= C_1x_1(t). \end{aligned}$$

It should be noted that RM2 is equivalent to applying RM1 to the dual (transposed) system of (1), followed by taking the transposed system again.

If $B \neq 0$ in RM1, differentiation of the difference part leads to derivatives of the input, whereas if $C \neq 0$ in RM2, the change of variable implies taking derivatives of the output. Derivatives of the input and output signals in the regularized system are not desired, and their absence is also of prime importance for the computation of $\|G\|_{\mathcal{H}_2}$, as shall become clear from Corollary 3.

The presented approach to handle the general case, where $B \neq 0$ and $C \neq 0$, consists of proving first that if the $\mathcal{H}_2$ norm is finite, system (1) has the same transfer matrix as the augmented system

$$\begin{aligned}
\frac{d}{dt}x_1(t) =& A_0^{(11)}x_1(t) + A_0^{(12)}x_2(t) + \sum_{j=1}^{m} A_j^{(11)}x_1(t-h_i) \\
&+ \sum_{j=1}^{m} A_j^{(12)}x_2(t-h_j) + B_1u(t) \\
x_2(t) =& A_0^{(21)}x_1(t) + \sum_{j=1}^{m} A_j^{(21)}x_1(t-h_j) \\
&+ \sum_{j=1}^{m} A_j x_2(t-h_j) + Bu(t) \\
x_3(t) =& A_0^{(21)}x_1(t) + \sum_{j=1}^{m} A_j^{(21)}x_1(t-h_j) \\
&+ \sum_{j=1}^{m} A_j x_3(t-h_j) \\
y(t) =& C_1x_1(t) + Cx_3(t),
\end{aligned} \tag{5}$$

where an additional variable x_3 is introduced, which allows to alternatively express the output equation (in terms of x_1 and x_3 instead of x_1 and x_2).

Next, the structure of (5) is such that we can apply RM1 and RM2 without differentiation of input or output. More precisely, since the equation for x_3 does not depend on the input we can turn it into a differential equation by applying RM1. Since the alternative output equation does not depend on x_2 we can apply RM2 and substitute x_2 by $\hat{\dot{x}}_2 + \hat{x}_2$. The additional dynamics, introduced by differentiation, are stable and either not controllable or not observable. In this way, we arrive at the following result.

Theorem 2. Let system (1) be strongly stable and condition (3)-(4) be satisfied. Consider the neutral type system

$$\begin{aligned}
\mathcal{D}_0\dot{x}(t) - \sum_{i=1}^{m} \mathcal{D}_i\dot{x}(t-h_i) =& \sum_{i=0}^{m} \mathcal{F}_i x(t-h_i) + \mathcal{B}u(t), \\
y(t) =& \mathcal{C}x(t),
\end{aligned} \tag{6}$$

where $x(t) \in \mathbb{R}^{2n+r}$ is a extended state vector composed by x_1, $\hat{x}_2$ and x_3, $\mathcal{B}^T := \begin{bmatrix} B_1^T & B^T & 0 \end{bmatrix}$, $\mathcal{C} := \begin{pmatrix} C_1 & 0 & C \end{pmatrix}$,

$$\mathcal{D}_0 = \begin{pmatrix} I & -A_0^{(12)} & 0 \\ 0 & I & 0 \\ -A_0^{(21)} & 0 & I \end{pmatrix}, \; \mathcal{D}_i = \begin{pmatrix} 0 & A_i^{(12)} & 0 \\ 0 & A_i & 0 \\ A_i^{(21)} & 0 & A_i \end{pmatrix}$$

and

$$\mathcal{F}_0 = \begin{pmatrix} A_0^{(11)} & A_0^{(12)} & 0 \\ A_0^{(21)} & -I & 0 \\ A_0^{(21)} & 0 & -I \end{pmatrix}, \; \mathcal{F}_i = \begin{pmatrix} A_i^{(11)} & A_i^{(12)} & 0 \\ A_i^{(21)} & A_i & 0 \\ A_i^{(21)} & 0 & A_i \end{pmatrix},$$

$i = 1, \ldots, m$, and its transfer matrix

$$\mathcal{G}(s) := \mathcal{C}\mathcal{H}^{-1}(s)\mathcal{B} \;\; s \in \mathbb{C} \setminus \Lambda_n$$

with $\Lambda_n := \{s \in \mathbb{C} : \det(\mathcal{H}(s)) = 0\}$, and

$$\mathcal{H}(s) = s\mathcal{D}_0 - s\sum_{i=1}^{m} \mathcal{D}_i e^{-sh_i} - \sum_{i=0}^{m} \mathcal{F}_i e^{-sh_i}.$$

It holds that

$$G(s) = \mathcal{G}(s), \; s \in \mathbb{C} \setminus \Lambda_n.$$

Moreover, the spectrum of system (6) satisfies

$$\Lambda_n = \Lambda \cup \Lambda_d \cup \{-1\}, \tag{7}$$

where $\Lambda_d := \{s \in \mathbb{C} : \det(A_{22}(s)) = 0\}$.

This result is particularly relevant for providing a formula for the computation of the $\mathcal{H}_2$ norm, as shown in the next corollary. Note that if the strong $\mathcal{H}_2$ norm is finite, it equals the $\mathcal{H}_2$ norm for the nominal delay values; see Theorem 1.

Corollary 3. If system (1) is strongly stable and condition (3)-(4) are satisfied, then we have

$$|||G|||_{\mathcal{H}_2} = \|G\|_{\mathcal{H}_2} = \|\mathcal{G}\|_{\mathcal{H}_2} = \sqrt{\mathrm{Tr}\,(\mathcal{B}^T U(0)\mathcal{B})},$$

where $U : [-h_m, h_m] \mapsto \mathbb{R}^{n\times n}$ is the so-called delay Lyapunov matrix associated with $\mathcal{C}^T\mathcal{C}$ of neutral type system (6).

3. FURTHER READING

For more details on the above notions and derivations and more illustrative examples, we refer to Gomez et al. (2020) and Mattenet et al. (2022).

REFERENCES

Gomez, M. A., Jungers, R. M., Michiels, W., 2020. On the m-dimensional Cayley–Hamilton theorem and its application to an algebraic decision problem inferred from the $\mathcal{H}_2$ norm analysis of delay systems. Automatica 113, 108761.

Mattenet, S., De Iuliis, V., Gomez, M. A., Michiels, W., Jungers, R. M., 2022. An improved finiteness test and a systematic procedure to compute the strong $\mathcal{H}_2$ norm of differential algebraic systems with multiple delays. Automatica, accepted for publication.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Hollow matrices and stabilization by noise

Tobias Damm * **Heike Faßbender** **

* *Department of Mathematics, TU Kaiserslautern, Germany, (e-mail damm@mathematik.uni-kl.de)*
** *Institute* Computational Mathematics*, TU Braunschweig, Germany, (e-mail h.fassbender@tu-bs.de)*

Abstract: We consider orthogonal transformations of arbitrary square matrices to a form where all diagonal entries are equal. In our main results we treat the simultaneous transformation of two matrices and the symplectic orthogonal transformation of one matrix. A relation to the joint real numerical range is worked out, efficient numerical algorithms are developed and applications to stabilization by rotation and by noise are presented.

Keywords: Matrix methods, simultaneous stabilization, stochastic systems, 15A21, 15B57, 93D15

1. INTRODUCTION

A square matrix whose diagonal entries are all zero, is sometimes called a *hollow matrix*, e.g. Charles et al. (2013); Farber and Johnson (2015); Kurata and Bapat (2016); Neven and Bastin (2018). By a theorem of Fillmore (1969), which is closely related to older results of Schur (1923); Horn (1954), every real square zero-trace matrix is orthogonally similar to a hollow matrix. Taken with a pinch of salt, the structure of a hollow matrix can be viewed as the negative of the spectral normal form (e.g. of a symmetric matrix), where the zeros are placed outside the diagonal. While the spectral form reveals an orthogonal basis of eigenvectors, a hollow form reveals an orthogonal basis of neutral vectors, i.e. vectors for which the quadratic form associated to the matrix vanishes.
This property turns out to be relevant in asymptotic eigenvalue considerations. More concretely, we use it to extend and give new proofs for results on stabilization of linear systems by rotational forces or by noise. Since the pioneering work of Arnold et al. (1983) these phenomena have received ongoing attention, with current interest e.g. in stochastic partial differential equations or Hamiltonian systems, Sri Namachchivaya and Vedula (2000); Caraballo and Robinson (2004); Kolba et al. (2019). Our new contribution concerns simultaneous stabilization by noise and features a new method of proof, which relies on an orthogonal transformation of matrices to hollow form.
It is easy to see that – in contrast to the spectral transformation – the transformation to hollow form leaves a lot of freedom to require further properties. In the present note, we first show that it is possible to transform two zero-trace matrices simultaneously to an almost hollow form, as will be specified in Section 2. In a non-constructive manner, the proof can be based on Brickman's theorem Brickman (1961) that the real joint numerical range of two real matrices is convex. Moreover, the simultaneous transformation result allows to prove a stronger version of Fillmore's theorem, namely that every real square zero-trace matrix is orthogonal-symplectically similar to a hollow matrix.
As an application, we show that a number of linear dissipative systems can be stabilized simultaneously by the same stochastic noise process, provided the coefficient matrices can be made almost hollow simultaneously by an orthogonal transformation.
The basis of this contribution is the paper Damm and Fassbender (2020), from which essential parts of this extended abstract are taken. But we also consider some more recent extensions and examples.

2. HOLLOW MATRICES AND ORTHOGONAL TRANSFORMATIONS

We first review some known facts on hollow matrices and then present our main results.

Definition 1. Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$.

(i) We call A *hollow*, if $a_{ii} = 0$ for all $i = 1, \dots, n$.
(ii) We call A *almost hollow*, if $a_{ii} = 0$ for $i = 3, \dots, n$ and $a_{11} = -a_{22}$.
(iii) We say that A is 2×2-*block hollow*, if $a_{ii} = -a_{i+1,i+1}$ for $i = 1, 3, \dots$ and $a_{nn} = 0$ in the case that n is odd.
(iv) If $\operatorname{trace} A = 0$, then A is called a *zero trace* matrix.

Obviously, 'hollow' $\Rightarrow$ 'almost hollow' $\Rightarrow$ '2×2-block hollow' $\Rightarrow$ 'zero trace'. Vice versa, $\operatorname{trace} A = 0$ implies that A is orthogonally similar to a hollow matrix. This result has been proven in Fillmore (1969). We add a proof, because similar arguments will be used later.

Lemma 2. Let $A \in \mathbb{R}^{n\times n}$ with $\operatorname{trace} A = 0$.

(a) There exists a vector $v \in \mathbb{R}^n$ with $v \neq 0$, such that $v^T A v = 0$.
(b) There exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$, such that $V^T A V$ is hollow.

Proof. (a) If $a_{11} = 0$, then we can choose $v = e_1$. Otherwise let (after possibly dividing A by a_{11}) w.l.o.g. $a_{11} = 1$. Since $\operatorname{trace} A = 0$, there exists $j \in \{2, \dots, n\}$ with $a_{jj} < 0$. For $v = xe_1 + e_j$ with $x \in \mathbb{R}$, we have

$$v^T A v = x^2 + (a_{1j} + a_{j1})x + a_{jj}$$

which has two real zeros. Hence (a) follows.
(b) Extend $v_1 = v/\|v\|$ with v from (a), to an orthonormal matrix $V_1 = [v_1, \ldots, v_n]$. Then $V^T A V = \begin{bmatrix} 0 & \star \\ \star & A_1 \end{bmatrix}$ with $A_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ and $\operatorname{trace} A_1 = \operatorname{trace} A = 0$. Therefore we can proceed with A_1 as with A.

Corollary 3. For $A \in \mathbb{R}^{n\times n}$, there exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$, such that all diagonal entries of $V^T A V$ are equal.

Proof. We set $A_0 = A - \frac{\operatorname{trace} A}{n} I$. By Lemma 2 there exists an orthogonal matrix V such that $V^T A_0 V$ is hollow. Then $V^T A V = V^T A_0 V + \frac{\operatorname{trace} A}{n} I$.

Remark 4. (a) A transformation matrix V making $V^T A V$ hollow as in Lemma 2 will sometimes be called an *(orthogonal) hollowiser (for A).*
(b) As is evident from the construction, the hollowiser V is not unique. In the following we will exploit this freedom to transform two matrices simultaneously or to replace V by an orthogonal symplectic matrix.
(c) Since $V^T A V$ is hollow, if and only if $V^T(A + A^T)V$ is hollow, there is no restriction in considering only symmetric matrices.
(d) We are mainly interested in the real case, but it is possible to transfer our results to the complex case, where $A \in \mathbb{C}^{n\times n}$ and V is unitary.

2.1 Simultaneous transformation of two matrices

Simultaneous transformation of several matrices to a certain form (e.g. spectral form) usually requires quite restrictive assumptions. Therefore it is remarkable that an arbitrary pair of zero trace matrices can simultaneously be transformed to an almost hollow pair. This is a consequence of a Brickman's theorem.

Theorem 5. (Brickman (1961)). Let $A, B \in \mathbb{R}^{n\times n}$ with $n \geq 3$. Then the set

$$W(A,B) = \{(x^T A x, x^T B x) \mid x \in \mathbb{R}^n, \|x\| = 1\}$$

is convex.

Proposition 6. Let $A, B \in \mathbb{R}^{n\times n}$ be zero trace matrices.

(a) If $n \geq 3$, there exists a nonzero vector $v \in \mathbb{R}^n$, such that $v^T A v = v^T B v = 0$.
(b) There exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ such that $V^T A V$ is hollow and $V^T B V$ is almost hollow.

Proof. (a): By Lemma 2, we can assume w.l.o.g. that A is hollow. If $b_{jj} = 0$ for some j, then we can choose $v = e_j$. Otherwise, since $\operatorname{trace} B = 0$, not all the signs of the b_{jj} are equal. For simplicity of notation assume that $b_{11} > 0$ and $b_{22} < 0$. The points $(e_1^T A e_1, e_1^T B e_1) = (0, b_{11})$ and $(e_2^T A e_2, e_2^T B e_2) = (0, b_{22})$ lie in the joint real numerical range of A and B, defined as

$$W(A,B) = \{(x^T A x, x^T B x) \mid x \in \mathbb{R}^n, \|x\| = 1\} \subset \mathbb{R}^2 .$$

According to Theorem 5 the set $W(A,B)$ is convex for $n \geq 3$. Hence it also contains $(0,0) = (v^T A v, v^T B v)$ for some unit vector $v \in \mathbb{R}^n$.
(b): Apply (a) repeatedly as in the proof of Lemma 2(b) until the remaining submatrix is smaller than 3×3 (where (a) is applied only for A).

Remark 7. The assumption in Proposition 6(a) that $n \geq 3$ is essential. As the standard counterexample consider $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. For $v = \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$, we have $v^T A v = x^2 - y^2$ and $v^T B v = 2xy$. If both forms are zero, then necessarily $x = y = 0$. Therefore, in general, a pair of symmetric matrices with zero trace is not simultaneously orthogonally similar to a pair of hollow matrices.

2.2 Symplectic transformation of a matrix

Symplectic transformations play an important role in Hamiltonian systems, e.g. Meyer et al. (2009). We briefly recapitulate some elementary facts. A real *Hamiltonian matrix* has the form

$$H = \begin{bmatrix} A & P \\ Q & -A^T \end{bmatrix} \in \mathbb{R}^{2n\times 2n} ,$$

where $A \in \mathbb{R}^{n\times n}$ is arbitrary, while $P, Q \in \mathbb{R}^{n\times n}$ are symmetric. If $J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$, then all real Hamiltonian matrices are characterized by the property that JH is symmetric. A real matrix $U \in \mathbb{R}^{2n\times 2n}$ is called *symplectic* if $U^T J U = J$. If U is symplectic, then the transformation $H \mapsto U^{-1} H U$ preserves the Hamiltonian structure. Amongst other things, symplectic orthogonal transformations are relevant for the Hamiltonian eigenvalue problem, e.g. Paige and van Loan (1981); van Loan (1984); Fassbender (2000). There is a rich theory on normal forms of Hamiltonian matrices under orthogonal symplectic transformations (e.g. Byers (1986); Lin et al. (1999)). It is, however, a surprising improvement of Lemma 2 that an arbitrary zero trace matrix can be hollowised by a symplectic orthogonal transformation.

Theorem 8. Consider a matrix $A \in \mathbb{R}^{2n\times 2n}$ with $n \geq 1$. Then there exists a symplectic orthogonal matrix U, such that $U^T A U$ has constant diagonal.

Proof. W.l.o.g. we can assume that A is symmetric with $\operatorname{trace} A = 0$. The transformation U is constructed in several steps, where we make use of the orthogonal symplectic transformations above.

1st step: Let $d_1, \ldots, d_{2n}$ denote the diagonal entries of A. Using a suitable symplectic Givens matrix $G_k(c,s)$ for the transformation $A^+ = G_k(c,s)^T A G_k(c,s)$ we can achieve that $d_k^+ = d_{k+n}^+$. After n such transformations we have

$$A^+ = \begin{bmatrix} A_1^+ & \star \\ \star & A_2^+ \end{bmatrix} = \begin{bmatrix} d_1^+ & & \star & & & \\ & \ddots & & \star & & \\ \star & & d_n^+ & & & \\ & & & d_1^+ & & \star \\ & \star & & & \ddots & \\ & & & \star & & d_n^+ \end{bmatrix} .$$

In particular $\operatorname{trace} A_1^+ = \operatorname{trace} A_2^+ = 0$.

2nd step: By Proposition 6, there exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$, such that $V^T A_1^+ V$ and $V^T A_2^+ V$ are (almost) hollow. Thus, for the symplectic orthogonal matrix $U = \begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix}$, we have (with $d_1 = 0$)

$$U^T A^+ U = \left[\begin{array}{cc|cc} \begin{bmatrix} d_1 & a \\ a & -d_1 \end{bmatrix} & \star & & \star \\ & \ddots & & \\ \star & 0 & & \\ \hline & & \begin{bmatrix} d_2 & b \\ b & -d_2 \end{bmatrix} & \star \\ \star & & & \ddots \\ & & \star & 0 \end{array}\right] .$$

3rd step: It suffices now to consider 4×4 matrices of the form $A = \begin{bmatrix} d_1 & a & \star & \star \\ a & -d_1 & \star & \star \\ \star & \star & d_2 & b \\ \star & \star & b & -d_2 \end{bmatrix}$. For these, a symplectic orthogonal transformation can be computed explicitly, but the details are omitted here.

3. APPLICATIONS TO STABILIZATION PROBLEMS

In this section we present two related stabilization problems. Both deal with unstable linear ordinary differential equations, whose coefficient matrix has negative trace. Such systems have stable and unstable modes, but the stable ones dominate. By a mixing of the modes the system can be stabilized. This mixing can be achieved e.g. by adding rotational forces or stochastic terms. For both cases we extend known results from the literature. The basic idea lies in an asymptotic analysis based on the hollow forms constructed in the previous sections.

3.1 Hamiltonian stabilization by rotation

A linear autonomous system $\dot{x} = Ax$ is called asymptotically stable, if all solutions $x(t)$ converge to 0 for $t \to \infty$. It is well known, that this is equivalent to the spectrum of A being contained in the open left half plane, $\sigma(A) \subset \mathbb{C}_-$. In this case, necessarily $\operatorname{trace} A < 0$. Vice versa, one can ask, whether for any matrix A with $\operatorname{trace} A < 0$, there exists a zero trace matrix M of a certain type, such that $\sigma(A + M) \subset \mathbb{C}_-$. In Crauel et al. (2007) it has been shown, that such a matrix M can always be chosen to be skew-symmetric. Then we say that M stabilizes A or by rotation, see e.g. Baxendale and Hennig (1993). The following theorem extends this result.

Theorem 9. Let $A \in \mathbb{R}^{2n\times 2n}$ with $\operatorname{trace} A < 0$. Then there exists a skew-symmetric Hamiltonian matrix M, such that $\sigma(A + M) \subset \mathbb{C}_-$.

3.2 Simultaneous stabilization by noise

Stabilization of a dynamic system by noise processes is an interesting phenomenon, which was analyzed in Arnold et al. (1983) (see also e.g. Arnold (1990); Caraballo and Robinson (2004)). As a particular situation, we consider the Stratonovich equation

$$dx = Ax\,dt + Mx \circ dw\ .$$

For a given matrix $A \in \mathbb{R}^{n\times n}$ we want to construct M such that this system is asymptotically mean square stable. It follows from results in Arnold et al. (1983) that this is possible (with a skew-symmetric M), if and only if $\operatorname{trace} A < 0$. Here we derive the following generalization.

Theorem 10. Let $A_1, A_2 \in \mathbb{R}^{n\times n}$ with $\operatorname{trace} A_1 < 0$ and $\operatorname{trace} A_2 < 0$ be given. Then there exists a common skew-symmetric matrix M, such that the systems

$$dx_1 = A_1 x_1\,dt + Mx_1 \circ dw_1$$
$$dx_2 = A_2 x_2\,dt + Mx_2 \circ dw_2$$

are both asymptotically mean square stable.

Remark 11. The proof of Theorem 10 exploits a simultaneous orthogonal transformation of $A_j - \frac{\operatorname{trace} A_j}{n} I$, $j = 1, 2$ to 2×2-block hollow form, which exists by Proposition 6. If, more generally, such a simultaneous transformation is possible for $m > 2$ matrices $A_j \in \mathbb{R}^{n\times n}$ with $\operatorname{trace} A_j < 0$ for $j = 1, \ldots, m$, then by the same argument there exists a common matrix M, such that all systems

$$dx_j = A_j x_j\,dt + Mx_j \circ dw_j \quad j = 1, \ldots, m$$

are asymptotically mean square stable.

4. OUTLOOK

Remark 11 raises the following question: How many zero-trace $n \times n$ matrices can be transformed simultaneously to 2×2-block hollow form by an orthogonal matrix V? Let us call such a set now briefly 2×2-hollowisable.

Note that the set of $n \times n$ orthogonal matrices constitutes a manifold of dimension $N = \frac{1}{2}n(n-1)$. We can view this as the number of degrees of freedom. To count the number of constraints let us assume n to be even. For a zero trace matrix $A \in \mathbb{R}^{n\times n}$, the condition that $\tilde{A} = V^T AV$ is 2×2-block hollow imposes the $C = \frac{1}{2}(n-1)$ conditions $\tilde{a}_{11} = -\tilde{a}_{22}, \tilde{a}_{33} = -\tilde{a}_{44}, \ldots, \tilde{a}_{n-3,n-3} = -\tilde{a}_{n-2,n-2}$. The last condition $\tilde{a}_{n-1,n-1} = -\tilde{a}_{n,n}$ is then satisfied automatically by the zero trace condition.
Thus, we might hope for $N/C = n$ zero trace $n \times n$ matrices to be 2×2-hollowisable. This is endorsed by numerical experiments with randomly generated zero trace matrices. On the other hand, the following lemmas provide counterexamples of, respectively, four 4×4 zero trace matrices and five zero trace matrices of arbitrary dimensions that are not 2×2-hollowisable. These results follow by elementary but tedious calculations, which are omitted.

Lemma 12. The set of matrices

$$\begin{bmatrix} 0&1&0&0\\1&0&0&0\\0&0&0&0\\0&0&0&0 \end{bmatrix}, \begin{bmatrix} 0&0&1&0\\0&0&0&0\\1&0&0&0\\0&0&0&0 \end{bmatrix}, \begin{bmatrix} 1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&-1 \end{bmatrix}, \begin{bmatrix} 1&0&0&1\\0&-1&0&0\\0&0&0&0\\1&0&0&0 \end{bmatrix} .$$

is not 2×2 hollowisable.

Lemma 13. For the matrices $B_1, \ldots, B_5$ given by

$$\begin{bmatrix} 1&0&0\\0&-1&0\\0&0&0 \end{bmatrix}, \begin{bmatrix} 1&0&0\\0&0&0\\0&0&-1 \end{bmatrix}, \begin{bmatrix} 0&1&0\\1&0&0\\0&0&0 \end{bmatrix}, \begin{bmatrix} 0&0&1\\0&0&0\\1&0&0 \end{bmatrix}, \begin{bmatrix} 0&0&0\\0&0&1\\0&1&0 \end{bmatrix}$$

let $A_j = \left[\begin{array}{c|c} B_j & 0 \\ \hline 0 & 0 \end{array}\right] \in \mathbb{R}^{n\times n}$ with $n \geq 3$ and $j = 1, \ldots, 5$.
Then the set $\{A_1, \ldots, A_5\}$ is not 2×2-hollowisable.

This gives rise to further connections and questions. For instance, a simultaneous transformation to a 2×2-block hollow form is related to the real 2-nd joint numerical range (cf. Fillmore and Williams (1971); Li and Poon (2000)). General conditions on the convexity of the real 2-nd numerical range (like e.g. in Gutkin et al. (2004)) do not seem to be available.
In particular, we have to discuss the following questions:

- Are three zero trace matrices always 2×2-hollowisable?
- Are four zero trace $n \times n$ matrices with $n > 4$ always 2×2-hollowisable?
- What general condition makes n zero trace $n \times n$ matrices 2×2-hollowisable?

REFERENCES

Arnold, L. (1990). Stabilization by noise revisited. *ZAMM Z. Angew. Math. Mech.*, 70(7), 235–246.

Arnold, L., Crauel, H., and Wihstutz, V. (1983). Stabilization of linear systems by noise. *SIAM J. Control Optim.*, 21, 451–461.

Baxendale, P. and Hennig, E. (1993). Stabilization of a linear system via rotational control. *Random Comput. Dyn.*, 1(4), 395–421.

Brickman, L. (1961). On the field of values of a matrix. *Proc. Amer. Math. Soc.*, 12, 61–66.

Byers, R. (1986). A Hamiltonian QR-algorithm. *SIAM J. Sci. Statist. Comput.*, 7, 212–229.

Caraballo, T. and Robinson, J.C. (2004). Stabilisation of linear PDEs by Stratonovich noise. *Syst. Control Lett.*, 53(1), 41–50.

Charles, Z.B., Farber, M., Johnson, C.R., and Kennedy-Shaffer, L. (2013). Nonpositive eigenvalues of hollow, symmetric, nonnegative matrices. *SIAM J. Matrix Anal. Appl.*, 34(3), 1384–1400.

Crauel, H., Damm, T., and Ilchmann, A. (2007). Stabilization of linear systems by rotation. *J. Differential Equations*, 234, 412–438.

Damm, T. and Fassbender, H. (2020). Simultaneous hollowisation, joint numerical range, and stabilization by noise. *SIAM J. Matrix Anal. Appl.*, 41(2), 637–656.

Farber, M. and Johnson, C.R. (2015). The structure of Schur complements in hollow, symmetric nonnegative matrices with two nonpositive eigenvalues. *Linear Multilinear Algebra*, 63(2), 423–438.

Fassbender, H. (2000). *Symplectic methods for the symplectic eigenproblem.* New York, NY: Kluwer Academic/Plenum Publishers.

Fillmore, P.A. and Williams, J.P. (1971). Some convexity theorems for matrices. *Glasg. Math. J.*, 12, 110–117.

Fillmore, P. (1969). On similarity and the diagonal of a matrix. *Amer. Math. Monthly*, 76(2), 167–169.

Gutkin, E., Jonckheere, E.A., and Karow, M. (2004). Convexity of the joint numerical range: Topological and differential geometric viewpoints. *Linear Algebra Appl.*, 376, 143–171.

Horn, A. (1954). Doubly stochastic matrices and the diagonal of a rotation matrix. *Am. J. Math.*, 76, 620–630.

Kolba, T., Coniglio, A., Sparks, S., and Weithers, D. (2019). Noise-induced stabilization of perturbed Hamiltonian systems. *Am. Math. Mon.*, 126(6), 505–518.

Kurata, H. and Bapat, R.B. (2016). Moore-Penrose inverse of a hollow symmetric matrix and a predistance matrix. *Spec. Matrices*, 4, 270–282.

Li, C.K. and Poon, Y.T. (2000). Convexity of the joint numerical range. *SIAM J. Matrix Anal. Appl.*, 21(2), 668–678.

Lin, W.W., Mehrmann, V., and Xu, H. (1999). Canonical forms for Hamiltonian and symplectic matrices and pencils. *Linear Algebra Appl.*, 302/303, 469–533.

Meyer, K.R., Hall, G.R., and Offin, D. (2009). *Introduction to Hamiltonian dynamical systems and the N-body problem.* Springer, New York, 2nd edition.

Neven, A. and Bastin, T. (2018). The quantum separability problem is a simultaneous hollowisation matrix analysis problem. *J. Phys. A*, 51(31).

Paige, C.C. and van Loan, C.F. (1981). A Schur decomposition for Hamiltonian matrices. *Linear Algebra Appl.*, 14, 11–32.

Schur, I. (1923). Über eine Klasse von Mittelbildungen mit Anwendungen auf die Determinantentheorie. Sitzungsber. Berl. Math. Ges. 22, 9-20.

Sri Namachchivaya, N. and Vedula, L. (2000). Stabilization of linear systems by noise: Application to flow induced oscillations. *Dyn. Stab. Syst.*, 15(2), 185–208.

van Loan, C.F. (1984). A symplectic method for approximating all the eigenvalues of a Hamiltonian matrix. *Linear Algebra Appl.*, 61, 233–251.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Discounted economic MPC without terminal conditions for periodic optimal behavior ⋆

Lukas Schwenkel * **Alexander Hadorn** * **Matthias A. Müller** **
Frank Allgöwer *

* *University of Stuttgart, Institute for Systems Theory and Automatic Control, 70550 Stuttgart, Germany (e-mail: {schwenkel@ist, st116411@stud, allgower@ist}.uni-stuttgart.de).*
** *Leibniz University Hannover, Institute of Automatic Control, 30167 Hannover, Germany, (e-mail: mueller@irt.uni-hannover.de)*

Abstract: In this work, we study economic model predictive control (MPC) in situations where the optimal operating behavior is periodic. In such a setting, the performance of a plain economic MPC scheme without terminal conditions can generally be far from optimal even with arbitrarily long prediction horizons. Whereas there are modified economic MPC schemes that guarantee optimal performance, all of them are based on prior knowledge of the optimal period length or of the optimal periodic orbit itself. In contrast to these approaches, we propose to achieve optimality by multiplying the stage cost by a linear discount factor. This modification is not only easy to implement but also independent of any system- or cost-specific properties, making the scheme robust against online changes therein. Under standard dissipativity and controllability assumptions, we can prove that the resulting linearly discounted economic MPC without terminal conditions achieves optimal asymptotic average performance up to an error that vanishes with growing prediction horizons. Moreover, we can guarantee practical asymptotic stability of the optimal periodic orbit under slightly stronger assumptions.

Keywords: Economic model predictive control, Optimal periodic operation, Turnpike property.
Mathematics Subject Classification (2020): Primary 93B45, 49N20.

1. INTRODUCTION

Economic model predictive control (MPC) (see, e.g., (Ellis et al., 2014), (Grüne and Pannek, 2017), (Faulwasser et al., 2018)) is an appealing control strategy for process control and other engineering applications due to its ability to directly optimize an economic criterion. In MPC, the control input is computed at each time step by solving a finite-horizon optimal control problem (OCP) online, in which the cost to be optimized can represent energy consumption, production amounts, or other physical or virtual costs. Whereas this control strategy is intuitive and in some examples very successful, its closed-loop performance can generally be far from optimal.

It is known that optimal operation at a steady state or at a periodic orbit is under a mild controllability condition equivalent to a certain dissipativity property (Müller et al., 2015). This property can be used to prove convergence of the closed loop to the optimal operating behavior by adding suitable terminal cost or terminal constraints to the OCP (see, e.g., (Amrit et al., 2011) for steady states, or (Zanon et al., 2017) for periodic orbits). Whereas this modification leads to an optimal asymptotic average performance of the closed loop, it requires significant offline design efforts including knowledge of a local control Lyapunov function with respect to the optimal steady state or optimal periodic orbit. A different approach is to implement the OCP without any terminal conditions. In case that steady-state operation is optimal, this plain MPC scheme achieves suboptimal asymptotic average performance under a similar dissipativity and controllability assumption, where the suboptimality is vanishing with growing prediction horizons (Grüne, 2013), (Grüne and Stieler, 2014). However, a fundamental limitation of this scheme is observed in (Müller and Grüne, 2016, Ex. 4): In case that periodic operation is optimal, the closed-loop asymptotic average performance can generally be far from optimal even for arbitrarily long prediction horizons. The problem is when the value function varies on the optimal periodic orbit, an unrewarding first step (e.g., waiting) may be preferred just to have a certain phase at the end of the prediction horizon. Since in MPC only the first step is actually implemented, it can cause a severe performance loss in closed loop if the first step is unrewarding. As solution, Müller and Grüne (2016) propose to implement a $p^\star$-step MPC scheme, where $p^\star$ is the optimal period length. Alternatively, Köhler et al. (2018) require the stage cost and the value function to be constant on the optimal periodic orbit. However, both solutions are not entirely satisfying since they either only work in a particular special case or still depend on the system and cost specific knowledge of the optimal period length such that an offline

⋆ F. Allgöwer and M. A. Müller are thankful that this work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – AL 316/12-2 and MU 3929/1-2 - 279734922. L. Schwenkel thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting him.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

design is necessary and needs to be repeated whenever the system or the economic cost change during operation.

In this work, we provide a solution to this latter problem and propose to mitigate the troubling effects at the end of the prediction horizon by multiplying the cost with a linear discount factor, which does not require any offline design and must not be adapted if the system or the cost change online. The main contribution of this work is a mathematical proof that this linearly discounted economic MPC scheme without terminal conditions achieves optimal asymptotic average performance up to an error vanishing with growing prediction horizons in situations where the optimal operating behavior is periodic. Moreover, we show that the MPC scheme renders the optimal orbit practically asymptotically stable. Both results are established based on a weaker version of the well known turnpike property, which is commonly used to analyze economic MPC without terminal conditions (e.g., (Grüne and Pannek, 2017)). However, due to space limitations, all proofs and numerical examples are omitted in this extended abstract and can be found in the journal version (Schwenkel et al., 2022).

The extended abstract is structured as follows: After denoting the problem setup more formally in Sec. 2 and defining the discounted OCP in Sec. 3, we state the main results in Sec. 4 and discuss our conclusions in Sec. 5.

Notation. We denote the set of naturals including 0 by $\mathbb{N}$, the set of reals by $\mathbb{R}$, and the set of integers in the interval $[a,b]$ by $\mathbb{I}_{[a,b]}$ for $a \leq b$ and define $\mathbb{I}_{[a,b]} = \emptyset$ for $a > b$. Further, we define the notation $[k]_p$ for the modulo operation, i.e., for the remainder when dividing k by p. Let $\mathcal{K}_\infty$ denote the set of continuous and monotonically increasing functions $\alpha : [0,\infty) \to [0,\infty)$ that satisfy $\alpha(0) = 0$ and $\lim_{t\to\infty} \alpha(t) = \infty$. Moreover, let $\mathcal{L}$ denote the set of continuous and monotonically decreasing functions $\delta : [0,\infty) \to [0,\infty)$ that satisfy $\lim_{t\to\infty} \delta(t) = 0$.

2. PROBLEM SETUP

In this section, we state the problem setup, which is to optimize the asymptotic average performance when controlling a nonlinear system that is optimally operated at a periodic orbit. As in (Müller and Grüne, 2016), (Zanon et al., 2017), and (Köhler et al., 2018), we consider a nonlinear discrete-time system

$$x(k+1) = f(x(k), u(k)) \tag{1}$$

subject to the constraints $x(k) \in \mathbb{X} \subset \mathbb{R}^n$ and $u(k) \in \mathbb{U} \subset \mathbb{R}^m$. We denote the trajectory resulting from a specific input sequence $u \in \mathbb{U}^N$ and the initial condition $x_0 \in \mathbb{X}$ with $x_u(k, x_0)$, defined by $x_u(0, x_0) = x_0$ and $x_u(k+1, x_0) = f(x_u(k, x_0), u(k))$ for $k \in \mathbb{I}_{[0,N-1]}$. Occasionally, we will use this notation also for feedback laws $\mu : \mathbb{X} \to \mathbb{U}$, in the natural sense $u(k) = \mu(x_\mu(k, x_0))$. Further, for each $x \in \mathbb{X}$ we denote the set of feasible control sequences of length N starting at x with $\mathbb{U}^N(x) := \{u \in \mathbb{U}^N | \forall k \in \mathbb{I}_{[0,N]} : x_u(k,x) \in \mathbb{X}\}$. The system is equipped with a stage cost function $\ell : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ and the control objective is to operate the system such that ℓ is minimized. To be more precise, for each $x \in \mathbb{X}$ and $u \in \mathbb{U}^T(x)$ we can define the accumulated cost

$$J_T(x,u) := \sum_{k=0}^{T-1} \ell(x_u(k,x), u(k)). \tag{2}$$

We are interested in finding a controller that generates in closed loop a sequence of inputs $u \in \mathbb{U}^\infty(x)$ such that the asymptotic average performance

$$J_\infty^{\mathrm{av}}(x,u) = \limsup_{T\to\infty} \frac{1}{T} J_T(x,u) \tag{3}$$

is minimized. All assumptions in this paper are equivalent to assumptions in Müller and Grüne (2016), especially the main results are stated under the same assumptions.

Assumption 1. (Continuity and compactness). *The functions f and ℓ are continuous, and the constraints $\mathbb{X} \times \mathbb{U}$ are compact. The set $\mathbb{X}$ is control invariant, i.e., $\mathbb{U}^\infty(x) \neq \emptyset$ for all $x \in \mathbb{X}$.*

The assumption of control invariance of $\mathbb{X}$ is commonly assumed in economic MPC without terminal conditions (e.g., in (Grüne, 2013), (Müller and Grüne, 2016), or (Köhler et al., 2018)) to simplify the analysis with the argument that it can be relaxed using methods similar to (Grüne and Pannek, 2017, Chapter 7.2 and 7.3).

Let us formally define (optimal, minimal) periodic orbits [1].

Definition 2. (Optimal periodic orbit). *A p-tuple $\Pi \in (\mathbb{X}\times\mathbb{U})^p$, $p \in \mathbb{N}$ is called a* feasible p-periodic orbit*, if its projection $\Pi_\mathbb{X}$ onto $\mathbb{X}^p$ satisfies*

$$\Pi_\mathbb{X}([k+1]_p) = f(\Pi(k)) \tag{4}$$

for all $k \in \mathbb{I}_{[0,p-1]}$. A p-periodic orbit Π is called minimal, if $\Pi_\mathbb{X}(k) = \Pi_\mathbb{X}(j) \Rightarrow k = j$ for all $k, j \in \mathbb{I}_{[0,p-1]}$. The distance of a pair $(x,u) \in \mathbb{X} \times \mathbb{U}$ to the orbit Π is defined as $\|(x,u)\|_\Pi := \inf_{k\in\mathbb{I}_{[0,p-1]}} \|(x,u) - \Pi(k)\|$. The set of all feasible p-periodic orbits is denoted by S_Π^p. The average cost at $\Pi \in S_\Pi^p$ is defined as $\ell^p(\Pi) := \frac{1}{p}\sum_{k=0}^{p-1} \ell(\Pi(k))$. If a feasible $p^\star$-periodic orbit $\Pi^\star$ satisfies

$$\ell^{p^\star}(\Pi^\star) = \inf_{p\in\mathbb{N},\, \Pi\in S_\Pi^p} \ell^p(\Pi) =: \ell^\star, \tag{5}$$

then $\Pi^\star$ is called an optimal periodic orbit *and $p^\star$ is called an* optimal period length*.*

Note that Ass. 1 guarantees that $\ell^\star$ in (5) is finite and that a minimizer $\Phi^\star$ exists. Further, note that in general there might be several (minimal) optimal orbits not only differing in their phase. However, if the following assumption of strict dissipativity (taken from (Köhler et al., 2018, Ass. 1)) is satisfied for a minimal orbit $\Pi^\star$, then this orbit is optimal and unique up to phase shifts. Further, strict dissipativity implies that the system is optimally operated at a periodic orbit, i.e., the best achievable asymptotic average performance is $\ell^\star$ (Müller et al., 2015).

Assumption 3. (Strict dissipativity). *There exists a storage function $\lambda : \mathbb{X} \to \mathbb{R}$ bounded on $\mathbb{X}$ and a function $\underline{\alpha}_{\tilde{\ell}} \in \mathcal{K}_\infty$, such that the rotated stage cost*

$$\tilde{\ell}(x,u) = \ell(x,u) - \ell^\star + \lambda(x) - \lambda(f(x,u)) \tag{6}$$

satisfies for all $x \in \mathbb{X}$ and all $u \in \mathbb{U}^1(x)$

$$\tilde{\ell}(x,u) \geq \underline{\alpha}_{\tilde{\ell}}(\|(x,u)\|_{\Pi^\star}). \tag{7}$$

[1] We use p-tuples $\Pi \in (\mathbb{X}\times\mathbb{U})^p$ as in (Zanon et al., 2017) instead of subsets $\Pi \subseteq \mathbb{X} \times \mathbb{U}$ with p elements as in (Müller and Grüne, 2016). The definition of *minimal* orbits, however, is analogous to (Müller and Grüne, 2016) to not only exclude multiple laps of the same orbit as in (Zanon et al., 2017) but to also exclude, e.g., 8-shaped orbits.

Additionally, we need the following two controllability conditions taken from (Müller and Grüne, 2016, Ass. 10 and 11).

Assumption 4. (Local controllability at $\Pi^\star$). *There exists $\kappa > 0, M' \in \mathbb{N}$ and $\rho \in \mathcal{K}_\infty$ such that for all $z \in \Pi^\star_{\mathbb{X}}$ and all $x, y \in \mathbb{X}$ with $\|x - z\| \leq \kappa$ and $\|y - z\| \leq \kappa$ there exists a control sequence $u \in \mathbb{U}^{M'}(x)$ such that $x_u(M', x) = y$ and*

$$\|(x_u(k,x), u(k))\|_{\Pi^\star} \leq \rho\big(\max\{\|x\|_{\Pi^\star_{\mathbb{X}}}, \|y\|_{\Pi^\star_{\mathbb{X}}}\}\big) \tag{8}$$

holds for all $k \in \mathbb{I}_{[0,M'-1]}$.

Assumption 5. (Finite-time reachability [2] of $\Pi^\star$). *For $\kappa > 0$ from Ass. 4 there exists $M'' \in \mathbb{N}$ such that for each $x \in \mathbb{X}$ there exists $K \in \mathbb{I}_{[0,M'']}$ and $u \in \mathbb{U}^K(x)$ such that $\|x_u(K,x)\|_{\Pi^\star_{\mathbb{X}}} \leq \kappa$.*

In (Köhler et al., 2018, Cor. 2) it is shown that the local controllability (Ass. 4) guarantees equivalence of the strict dissipativity assumptions from our setup (Ass. 3) and from (Müller and Grüne, 2016, Ass. 9). Hence, we impose equivalent assumptions as Müller and Grüne (2016).

3. LINEARLY DISCOUNTED MPC SCHEME

In this section, we define the linearly discounted economic MPC scheme starting with the finite-horizon discounted cost functional

$$J_N^\beta(x,u) := \sum_{k=0}^{N-1} \beta_N(k)\ell(x_u(k,x), u(k)) \tag{9}$$

with the linear discount function $\beta_N(k) := \frac{N-k}{N}$. Further, the corresponding optimal value function is

$$V_N^\beta(x) := \inf_{u \in \mathbb{U}^N(x)} J_N^\beta(x,u). \tag{10}$$

Due to Ass. 1 we know that J_N^β is continuous and that for each $x \in \mathbb{X}$ the set $\mathbb{U}^N(x)$ is nonempty and compact. Therefore, there exists for each $x \in \mathbb{X}$ a possibly non-unique input sequence $u^\beta_{N,x} \in \mathbb{U}^N(x)$ that attains the infimum, i.e., $V_N^\beta(x) = J_N^\beta(x, u^\beta_{N,x})$. Then we can define the standard MPC feedback law

$$\mu_N^\beta(x) := u^\beta_{N,x}(0), \tag{11}$$

that is, for a given x we minimize $J_N^\beta(x,u)$ and take the first element of the, or if non-unique of some, minimizing sequence $u^\beta_{N,x}$.

Before we start analyzing the closed-loop performance of this scheme, we want to share some intuition how discounting can be beneficial when dealing with a periodic optimal behavior. In the counter example in (Müller and Grüne, 2016, Exmp. 4) a plain economic MPC scheme without terminal conditions fails to converge to the optimal periodic orbit for all (arbitrarily long) odd prediction horizons. The problem occurring therein is that a certain phase at the end of the prediction horizon is preferred such that for any odd prediction horizon, it is better to first wait one time step before approaching the 2-periodic orbit. In closed loop, the control law (11) leads to waiting forever. A discount can overcome this problem, as the reward of ending at the right phase is discounted more than the cost of waiting at the first time step.

The idea of discounting the stage cost function is not new, however, in the context of economic MPC there are only a few works considering exponential discounts, such as, for example, (Grüne et al., 2016), (Grüne et al., 2021), (Zanon and Gros, 2022). In our setup, an exponential discount $\beta_N(k) = \beta^k$ for some $\beta \in (0,1)$ would decrease too fast as we need $\lim_{N\to\infty} \sum_{k=0}^{N-1} \beta_N(k) = \infty$ to make sure that the reward in the discounted cost function of being at the optimal orbit is larger than any transient cost of approaching it as long as the prediction horizon is sufficiently large. As the following section shows, a linear discount factor provides not only this property but is also appealing to analyze since we can exploit the linearity. Linear discounts are much less common than exponential discounts, nonetheless, a similar linear discount factor has also been used by Soloperto et al. (2022), however, in a different context of learning-based MPC.

4. OPTIMAL AVERAGE PERFORMANCE

In this section, we discuss the key contribution of this paper: The linearly discounted economic MPC schemes without terminal conditions from Sec. 3 achieves an asymptotic average performance that is optimal up to an error vanishing with growing prediction horizons. This performance result is analogous to the results known from other economic MPC schemes without terminal conditions, compare (Grüne and Stieler, 2014) in case of optimal steady-state operation or (Müller and Grüne, 2016) in case of optimal periodic operation using a p-step MCP scheme. In these works, the proof of the performance bound is heavily based on the so-called *turnpike-property*, which states that solutions of the OCP stay for all but a fixed number (independent of the length of the prediction horizon) of time steps in the neighborhood of the optimal behavior. Unfortunately, when discounting the stage cost we jeopardize this property as due to the small weights at the end of the horizon, more and more points could lie outside the neighborhood, hence, this number now depends on the length of the prediction horizon. The following theorem shows that the number of points in the neighborhood still grows faster than the number of points outside, which we therefore call the *weak turnpike property*.

Theorem 6. (Weak turnpike property). *Let Ass. 1, 3, 4, and 5 hold. For $\varepsilon > 0$, $N \in \mathbb{N}$ and $x \in \mathbb{X}$, define the number of points of the optimal trajectory $u^\beta_{N,x}$ in an ε-neighborhood of the optimal orbit $\Pi^\star$ as*

$$Q_\varepsilon^\beta(N,x) = \# \left\{ k \in \mathbb{I}_{[0,N-1]} \;\middle|\; \left\| \left(x_{u^\beta_{N,x}}(k,x), u^\beta_{N,x}(k)\right) \right\|_{\Pi^\star} \leq \varepsilon \right\}. \tag{12}$$

Then, there exist $\alpha_1 \in \mathcal{K}_\infty$ such that

$$Q_\varepsilon^\beta(N,x) \geq N - \frac{\sqrt{N}}{\alpha_1(\varepsilon)} \tag{13}$$

holds for all $x \in \mathbb{X}$, $N \in \mathbb{N}$, and $\varepsilon > 0$.

Remember that in the commonly known turnpike property (see, e.g., (Grüne and Pannek, 2017)) the $\sqrt{N}$ term in (13) is a constant independent of N. Hence, whereas the

[2] Technically, Ass. 5 guarantees finite-time reachability of a neighborhood of $\Pi^\star$. Together with the local controllability Ass. 4, reachability of $\Pi^\star$ in $M'' + M'$ steps follows.

weak turnpike property does not imply that the number of points outside the ε-neighborhood $N - Q_\varepsilon^\beta(N,x)$ is bounded by a constant, it still satisfies that the proportion of points inside is growing to 1, i.e., $\lim_{N\to\infty} \frac{1}{N} Q_\varepsilon^\beta(N,x) = 1$. This weaker property is actually still enough to show the following performance bound, which is analogous to the one from (Müller and Grüne, 2016).

Theorem 7. (Asymptotic average performance). *Let Ass. 1, 3, 4, and 5 hold. Then there exists $\delta \in \mathcal{L}$ such that for each prediction horizon length $N \in \mathbb{N}$, the MPC feedback law μ_N^β defined in* (11) *results in an asymptotic average performance that is not worse than*

$$J_\infty^{\mathrm{av}}(x, \mu_N^\beta) \leq \ell^\star + \delta(N). \tag{14}$$

If the storage function λ is continuous and the optimal orbit $\Pi^\star$ is minimal, we can even guarantee practical asymptotic stability of the optimal periodic orbit $\Pi^\star$.

Theorem 8. (Practical asymptotic stability). *Let Ass. 1, 3, 4, and 5 hold with a continuous storage function λ and assume that Π^* is a minimal orbit. Then there exists $\varepsilon \in \mathcal{L}$ such that the optimal periodic orbit $\Pi^\star$ is practically asymptotically stable under the MPC feedback μ_N^β w.r.t. $\varepsilon(N)$, i.e., there exists $\beta \in \mathcal{KL}$ and $l \in \mathbb{I}_{[0,p-1]}$ such that for all $x \in \mathbb{X}$ and $k \in \mathbb{N}$ we have*

$$\|x_\mu(k,x) - \Pi_{\mathbb{X}}([k+l]_p)\| \leq \max\{\beta(\|x\|_{\Pi_{\mathbb{X}}}, k), \varepsilon(N)\}. \tag{15}$$

5. DISCUSSION AND CONCLUSION

In this work, we have shown that a linearly discounted economic MPC scheme without terminal conditions achieves an asymptotic average performance that is optimal up to any desired level of suboptimality when the prediciton horizon is sufficiently large. The main novelty of this work is that both the proposed scheme and the performance guarantee are independent of the optimal period length $p^\star$ (compared to (Grüne, 2013), which only holds for $p^\star = 1$, i.e., steady states, and (Müller and Grüne, 2016), which uses a $p^\star$-step MPC scheme).

When facing real world applications, it is in most cases very difficult or even impracticable to design terminal conditions. Often, the only practicable solution is to omit terminal conditions and increase the prediction horizon N until the closed-loop behavior is satisfactory. The work of Grüne (2013) provides a theoretical justification for this procedure in the case where optimal operation is a steady state. Similarly, in the case where optimal operation is a steady state *or* a periodic orbit, the present work provides a theoretical justification to implement a linearly discounted economic MPC schemes without terminal conditions and increase its prediction horizon N until the desired performance is reached.

Future work will need to investigate what length of the prediction horizon is actually needed for a decent performance. We expect that the stronger theoretical guarantees come at a price. In cases where both discounted and non-discounted schemes guarantee near-optimal asymptotic average performance (e.g., when the optimal behavior is a steady state) the discount might lead to an increase in the suboptimality bound $\delta(N)$. On the other hand, having practical asymptotic stability (Theorem 8) is expected to result in a better transient performance compared to having only practical convergence as in the $p^\star$-step MPC scheme from (Müller and Grüne, 2016). This expectation can be justified with (Grüne and Stieler, 2014), where transient performance guarantees are based on practical asymptotic stability. The simulation study in (Schwenkel et al., 2022) sheds more light on the differences between the schemes and supports our expectations. Moreover, the results of this work suggest that also other discount functions may guarantee the same qualitative performance bound, which could lead to interesting insights how different discounts affect the performance.

REFERENCES

Amrit, R., Rawlings, J.B., and Angeli, D. (2011). Economic optimization using model predictive control with a terminal cost. *Annual Reviews in Control*, 35(2), 178 – 186. doi: 10.1016/j.arcontrol.2011.10.011.

Ellis, M., Durand, H., and Christofides, P.D. (2014). A tutorial review of economic model predictive control methods. *J. Process Control*, 24(8), 1156–1178. doi:10.1016/j.jprocont.2014.03.010.

Faulwasser, T., Grüne, L., and Müller, M.A. (2018). Economic nonlinear model predictive control. *Foundations and Trends in Systems and Control*, 5, 1–98. doi:10.1561/2600000014.

Grüne, L. (2013). Economic receding horizon control without terminal constraints. *Automatica*, 49(3), 725 – 734. doi: 10.1016/j.automatica.2012.12.003.

Grüne, L. and Stieler, M. (2014). Asymptotic stability and transient optimality of economic MPC without terminal conditions. *J. Process Control*, 24(8), 1187 – 1196. doi: 10.1016/j.jprocont.2014.05.003.

Grüne, L., Kellett, C.M., and Weller, S.R. (2016). On a discounted notion of strict dissipativity. In *IFAC Symp. Nonlinear Control Systems (NOLCOS)*, volume 49, 247–252. Monterey, CA, USA. doi:10.1016/j.ifacol.2016.10.171.

Grüne, L., Müller, M.A., Kellett, C.M., and Weller, S.R. (2021). Strict dissipativity for discrete time discounted optimal control problems. *Mathematical Control & Related Fields*, 11(4), 771. doi:10.3934/mcrf.2020046.

Grüne, L. and Pannek, J. (2017). *Nonlinear Model Predictive Control: Theory and Algorithms*. Springer, Cham, Switzerland. doi:10.1007/978-3-319-46024-6.

Köhler, J., Müller, M.A., and Allgöwer, F. (2018). On periodic dissipativity notions in economic model predictive control. *IEEE Control Systems Letters*, 2(3), 501–506. doi: 10.1109/lcsys.2018.2842426.

Müller, M.A. and Grüne, L. (2016). Economic model predictive control without terminal constraints for optimal periodic behavior. *Automatica*, 70, 128–139. doi:10.1016/j.automatica.2016.03.024.

Müller, M.A., Grüne, L., and Allgöwer, F. (2015). On the role of dissipativity in economic model predictive control. In *5th IFAC Conf. Nonlinear Model Predictive Control (NMPC)*, 23, 110–116. Seville, Spain. doi:10.1016/j.ifacol.2015.11.269.

Schwenkel, L., Hadorn, A., Müller, M.A., and Allgöwer, F. (2022). Linearly discounted economic MPC without terminal conditions for periodic optimal operation. *Currently under review, preprint available on arXiv:* 2205.03118.

Soloperto, R., Müller, M.A., and Allgöwer, F. (2022). Guaranteed closed-loop learning in model predictive control. *IEEE Trans. Automat. Control*. doi:10.1109/TAC.2022.3172453.

Zanon, M. and Gros, S. (2022). A new dissipativity condition for asymptotic stability of discounted economic MPC. *Automatica*, 141, 110287. doi:10.1016/j.automatica.2022.110287.

Zanon, M., Grüne, L., and Diehl, M. (2017). Periodic optimal control, dissipativity and MPC. *IEEE Trans. Autom. Control*, 62(6), 2943–2949. doi:10.1109/tac.2016.2601881.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Towards a Stochastic Fundamental Lemma for LTI Systems

Timm Faulwasser, Guanru Pan, and Ruchuan Ou

Institute for Energy Systems, Energy Efficiency and Energy Economics, TU Dortmund University, Dortmund, Germany (e-mail: timm.faulwasser@ieee.org).

Abstract: Jan Willems and co-authors introduced the characterization of finite-time behaviors of linear systems via the image of Hankel matrices already in 2005. The increasing popularity and research interest of data-driven control techniques has catalyzed the use of this result—which is commonly known as Willems' fundamental lemma—for predictive control and beyond. In this note, we recap recent results on a stochastic extension of the fundamental lemma from (Pan et al., 2021). Specifically, we leverage the framework of polynomial chaos expansions to derive a computationally tractable stochastic extension of the fundamental lemma.

Keywords: Stochastic systems, Hankel matrices, data-driven control, polynomial chaos expansions

1. INTRODUCTION

Data-driven system descriptions based on the fundamental lemma by Willems et al. (2005) are subject to continued research interest. Essentially, the lemma states that the trajectories of any controllable Linear Time Invariant (LTI) system can be represented without explicit knowledge of a state-space model. Specifically, provided persistency of excitation holds, over any finite horizon the manifest system behavior is contained in the column space of a Hankel matrix constructed from recorded input-output data. In absence of process and measurement noise, this representation is exact. There are recent variants of the original result, e.g., extensions to nonlinear systems (Alsalti et al., 2021; Lian et al., 2021), to linear parameter-varying and time-varying systems (Verhoek et al., 2021; Nortmann and Mylvaganam, 2020), to linear network systems (Allibhoy and Cortés, 2020), and to linear descriptor systems (Schmitz et al., 2022). We refer to De Persis and Tesi (2019); Markovsky and Dörfler (2021) for recent overviews. Data-driven control design and system analysis with not necessarily persistently exciting data has been investigated by van Waarde et al. (2020b).

Beyond the LTI setting, Coulson et al. (2019) propose a heuristic approach to deal with measurement noise and mild system nonlinearities by introducing slack variables and regularization in the objective function. There is also a line of research focusing on the robustness with respect to measurement noise and/or process noise. While Berberich et al. (2020); De Persis and Tesi (2021); van Waarde et al. (2020a) consider the design of robust state feedback controllers to deal with process noise, Yin et al. (2020) uses maximum likelihood concepts to obtain an optimal Hankel representation, and Coulson et al. (2021) views the noise entering the Hankel matrix as a problem of distributional robustness.

However, to the best of the authors' knowledge, so far there appears to be no stochastic variant of the fundamental lemma. Hence, this note recalls the main results of a recent submission (Pan et al., 2021), wherein we presented a stochastic fundamental lemma for LTI systems. To this end, we rely on Polynomial Chaos Expansion (PCE) which is an established method that can be applied in Markovian and non-Markovian settings. Its core idea is based on the observation that random variables can be regarded as elements of an $\mathcal{L}^2$ probability space and hence they admit representations in appropriately chosen polynomial bases (Sullivan, 2015).

This note recaps main results of Pan et al. (2021). The remainder is structured as follows: In Section 2 we recall the problem statements, while in Section 3 we present selected results without proofs. Finally, the note ends with conclusions in Section 4.

2. PROBLEM STATEMENT

We consider LTI systems of the following form

$$x_{k+1} = Ax_k + Bu_k + Ew_k, \quad x_0 = x_{\text{init}} \in \mathbb{R}^{n_x} \tag{1a}$$

$$y_k = Cx_k + Du_k \tag{1b}$$

where $x \in \mathbb{R}^{n_x}$, $u \in \mathbb{R}^{n_u}$, $w \in \mathbb{R}^{n_u}$, and $y \in \mathbb{R}^{n_y}$ refer to the state, the input, the disturbance, and the output.

Definition 1. (Persistency of excitation). Let $T, t \in \mathbb{N}^+$. A sequence of inputs $\mathbf{u}_{[0,T-1]}$ is said to be persistently exciting of order t if the Hankel matrix

$$\mathcal{H}_t(\mathbf{u}_{[0,T-1]}) \doteq \begin{bmatrix} u_0 & u_1 & \cdots & u_{T-t} \\ u_1 & u_2 & \cdots & u_{T-t+1} \\ \vdots & \vdots & \ddots & \vdots \\ u_{t-1} & u_t & \cdots & u_{T-1} \end{bmatrix}$$

is of full row rank. □

⋆ RO and TF acknowledge funding by the German Federal Ministry of Education and Research (BMBF) in the course of the 6GEM research hub under grant number 16KISK038.

Moreover, since (1) is driven by the inputs and the process noise realizations, the extension of the fundamental lemma of Willems et al. (2005) to the exogenous input data $(\mathbf{u}, \mathbf{w})_{[0,T-1]}$ is immediate. Below we use the notation $\forall \mathbf{z} \in \{\mathbf{x}, \mathbf{u}, \mathbf{w}, \mathbf{y}\}$ to highlight equations which have to hold for a set of variables.

Lemma 1. (Deterministic fundamental lemma). Let $T, t \in \mathbb{N}^+$. If $(\mathbf{u}, \mathbf{w})_{[0,T-1]}$ is persistently exciting of order $n_x + t$ and (1) is controllable with respect to the exogenous inputs (u, w), then $(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{w}}, \tilde{\mathbf{y}})_{[0,t-1]}$ is an element of the behavior of (1) if and only if there exists a $g \in \mathbb{R}^{T-t+1}$ such that

$$\mathcal{H}_t(\mathbf{z}_{[0,T-1]})g = \tilde{\mathbf{z}}_{[0,t-1]}, \quad \forall \mathbf{z} \in \{\mathbf{x}, \mathbf{u}, \mathbf{w}, \mathbf{y}\}. \qquad \square$$

The above result provides a non-parametric system description of the LTI system (1). It allows to capture the dynamics, once the uncertainty surrounding the disturbance w has realized (or when it is sampled). Hence, we refer to (1) as *realization dynamics*. The crux is, however, that the future evolution of the disturbance $w \in \mathbb{R}^{n_w}$ is usually not known or difficult to predict.

Thus, an alternative modelling can be done in terms of random variables. This yields

$$X_{k+1} = AX_k + BU_k + EW_k, \quad X_0 = X_{\text{init}} \tag{2a}$$
$$Y_k = CX_k + DU_k, \tag{2b}$$

with state $X_k \in \mathcal{L}^2(\Omega, \mathcal{F}_k, \mu; \mathbb{R}^{n_x})$, where Ω is the sample space, $\mathcal{F}$ is a σ-algebra, $\mathbb{F} \doteq (\mathcal{F}_k)_{k=0,\dots,N}$ is a stochastic filtration, and μ is the considered probability measure. In the underlying filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mu)$, the σ-algebra contains all available historical information. Likewise, the stochastic input U_k is modelled as a stochastic process that is adapted to the filtration $\mathbb{F}$, that is, U_k only depends on $X_0, X_1, ..., X_k$. Note that the influence of the noise $W_i, i \leq k$ is implicitly handled via the state recursion. For more details on filtrations we refer to Fristedt and Gray (2013).

The process noise W_k, $k \in \mathbb{N}$ is considered as *i.i.d.* random variables whose underlying probability distribution is assumed to be known. Additionally, the distribution of the initial condition X_{init} is also assumed to be known. We remark that neither the distribution of X_{init} nor the one of W_k, $k \in \mathbb{N}$ needs to be Gaussian. Indeed, under mild assumptions, our proposed framework is applicable to random variables of finite variance.

3. SELECTED RESULTS

Considering the stochastic system (2), we are interested in deriving a counterpart to Lemma 1. To this end, we recap the basics of Polynomial Chaos Expansion (PCE).

3.1 Basics of Polynomial Chaos Expansion

PCE enables the propagation of uncertainties through dynamics and it provides a finite dimensional representation of random variables. Its origins date back to Wiener (1938); for a general introduction to PCE see Sullivan (2015).

The main idea of PCE is that an $\mathcal{L}^2$ random variable can be expressed in a suitable polynomial basis. To this end, an orthogonal polynomial basis $\{\phi^j(\omega)\}_{j=0}^{\infty}$ which spans $\mathcal{L}^2(\Omega, \mathcal{F}, \mu; \mathbb{R})$, i.e.,

$$\langle \phi^i, \phi^j \rangle \doteq \int_\Omega \phi^i(\omega)\phi^j(\omega)\,\mathrm{d}\mu(\omega) = \delta^{ij}\langle \phi^j, \phi^j \rangle,$$

where δ^{ij} is the Kronecker delta, is considered. The PCE of a real-valued random variable $V \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu; \mathbb{R})$ with respect to this basis is given by

$$V = \sum_{j=0}^{\infty} \mathsf{v}^j \phi^j \text{ with } \mathsf{v}^j = \frac{\langle V, \phi^j \rangle}{\langle \phi^j, \phi^j \rangle}, \tag{3}$$

where $\mathsf{v}^j \in \mathbb{R}$ is called the j-th PCE coefficient.

Applying PCE component-wise the j-th PCE coefficient of a vector-valued random variable $V \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu; \mathbb{R}^n)$ is

$$\mathsf{v}^j = \begin{bmatrix} \mathsf{v}^{1,j} & \mathsf{v}^{2,j} & \dots & \mathsf{v}^{n,j} \end{bmatrix}^\top,$$

where $\mathsf{v}^{i,j}$ is the j-th PCE coefficient of component V^i.

In numerical implementations the series has to be terminated after a finite number of terms, i.e., one works with

$$V = \sum_{j=0}^{L-1} \mathsf{v}^j \phi^j, \tag{4}$$

where $L \in \mathbb{N}$. Naturally, this may lead to truncation errors. For the purpose of this note, we assume that all considered PCE series are exact with finite expansion order $L-1$ (in the $\mathcal{L}^2$-equivalence sense). For a detailed investigation of how exactness of the uncertainty propagation for the LTI system (2) we refer to Mühlpfordt et al. (2018).

Remark 1. (Cond. probabilities, moments, and PCEs). From a stochastic control perspective, it is natural to ask how the PCE framework sketched above relates to widely used concepts such as conditional probabilities, conditional probability densities, and statistical moments. For starters, we remark that the moments are obtained as polynomial functions of the PCE coefficients (Sullivan, 2015). Due to the fact that the $\mathcal{L}^2$ framework covers a wide range of random variables, it is not immediate to give closed-form expressions of probability densities in the PCE framework. Yet, given a PCE for $Z = f(X, C)$ where $X, C \in \mathcal{L}^2(\Omega, \mathcal{F}_k, \mu; \mathbb{R}^{n_x})$ are independently distributed, conditional densities and probabilities (e.g. $P(Z \,|\, C = c)$ can be approximated via sampling of $Z = f(X, c)$ over X. While the numerical details are skipped due to space limitations, it is crucial to note that density information and conditional probability information is not lost using PCEs. $\square$

3.2 Stochastic Fundamental Lemmata

Replacing all random variables of (2) with corresponding PCE expansions with respect to the basis $\{\phi^j(\omega)\}_{j=0}^{\infty}$ and performing a Galerkin projection onto the basis functions $\phi^j(\omega)$, one obtains the dynamics of the PCE coefficients with given initial condition $\mathsf{x}^j_{\text{init}}$ for $j \in \{0, \dots, L-1\}$

$$\mathsf{x}^j_{k+1} = A\mathsf{x}^j_k + B\mathsf{u}^j_k + E\mathsf{w}^j_k, \quad \mathsf{x}^j_0 = \mathsf{x}^j_{\text{init}}, \tag{5a}$$
$$\mathsf{y}^j_{k+1} = C\mathsf{x}^j_k + D\mathsf{u}^j_k, \quad \forall j \in \{0, \dots, L-1\} \tag{5b}$$

Since (5) is a deterministic LTI system, it allows the conceptual application of the usual LTI fundamental lemma.

Lemma 2. (Fundamental lemma for PCE coefficients). Let $T, t \in \mathbb{N}^+$. For all $j \in \{0, \dots, L-1\}$ and given a trajectory tuple $(\mathsf{x}, \mathsf{u}, \mathsf{w}, \mathsf{y})^j_{[0,T-1]}$ generated by (5), suppose $(\mathsf{u}, \mathsf{w})^j_{[0,T-1]}$ is persistently exciting of order $n_x + t$

and (5) is controllable with respect to $(\mathsf{u}^j, \mathsf{w}^j)$. Then, for all $j \in \{0, \dots, L-1\}$, $(\tilde{\mathsf{x}}, \tilde{\mathsf{u}}, \tilde{\mathsf{w}}, \tilde{\mathsf{y}})^j_{[0,t-1]}$ is an element of the behavior of (5) if and only if there exists a $\mathsf{g}^j \in \mathbb{R}^{T-t+1}$ such that

$$\mathcal{H}_t(\mathsf{z}^j_{[0,T-1]})\mathsf{g}^j = \tilde{\mathsf{z}}^j_{[0,t-1]}, \quad \forall \mathsf{z} \in \{\mathsf{x}, \mathsf{u}, \mathsf{w}, \mathsf{y}\}. \qquad \square$$

The proof follows directly from the results of Willems et al. (2005). Lemma 2 as such is straightforward but of limited practical use. Consider the case that the disturbance w is i.i.d., i.e. identically and independently distributed. In terms of PCE this implies that the coefficients modelling W_k, w^j_k, are constant for all k. Hence, persistency of excitation cannot be satisfied. Moreover, it is not trivial to measure/estimate PCE coefficients of a stochastic LTI system from data.

Recently, we made an observation which allows to overcome the issue of persistency of excitation in the PCE coefficient dynamics (5) (Pan et al., 2021). Specifically, it is worth to be noted—in a model-based setting—that the realization dynamics (1), the stochastic dynamics (2), and the dynamics of the PCE coefficients (5) are all parametrized by the same matrices (A, B, C, D, E). This, in turn, implies that the subspaces spanned by available data are equivalent (provided persistency of excitation holds).

To the end of exploiting this observation, we introduce the notation $(\mathbf{z}, \mathbf{Z}) \in \{(\mathbf{x}, \mathbf{X}), (\mathbf{u}, \mathbf{U}), (\mathbf{w}, \mathbf{W}), (\mathbf{y}, \mathbf{Y})\}$ which expresses that $\mathbf{x}$ corresponds as a realization trajectory to the random-variable trajectory $\mathbf{X}$ and likewise for $(\mathbf{u}, \mathbf{U}), (\mathbf{w}, \mathbf{W}), (\mathbf{y}, \mathbf{Y})$.

Lemma 3. (Column-space equivalence). Consider system (2) and an $\mathcal{L}^2(\Omega, \mathcal{F}_k, \mu; \mathbb{R}^{n_z})$, $n_z \in \{n_x, n_u, n_x, n_y\}$ random-variable trajectory tuple $(\mathbf{X}, \mathbf{U}, \mathbf{W}, \mathbf{Y})_{[0,T-1]}$. Let the corresponding PCE coefficient trajectories $(\mathsf{u}, \mathsf{w})^j_{[0,T-1]}$, $j \in \{0, \dots, L-1\}$ and the realizations $(\mathbf{u}, \mathbf{w})_{[0,T-1]}$ be persistently exciting of order n_x+t, and let (1) be controllable with respect to (u, w).

(i) Then, for all $j \in \{0, \dots, L-1\}$ and all $(\mathbf{z}, z) \in \{(\mathbf{x}, \mathsf{x}), (\mathbf{u}, \mathsf{u}), (\mathbf{w}, \mathsf{w}), (\mathbf{y}, \mathsf{y})\}$, it holds that

$$\operatorname{colsp}\big(\mathcal{H}_t(\mathsf{z}^j_{[0,T-1]})\big) = \operatorname{colsp}\big(\mathcal{H}_t(\mathbf{z}_{[0,T-1]})\big). \qquad (6a)$$

(ii) Moreover, for all $g \in \mathbb{R}^{T-t+1}$, there exists a $G \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu; \mathbb{R}^{T-t+1})$ such that

$$\mathcal{H}_t(\mathbf{Z}_{[0,T-1]})g = \mathcal{H}_t(\mathbf{z}_{[0,T-1]})G. \qquad (6b)$$

for all $(\mathbf{z}, Z) \in \{(\mathbf{x}, X), (\mathbf{u}, U), (\mathbf{w}, W), (\mathbf{y}, Y)\}$. $\square$

In slightly different form this result has appeared as Lemma 3 in Pan et al. (2021). Hence its proof is omitted.

The crucial observation is that in (6a) PCE data appears in the Hankel matrices on the left hand side, while realization data is used on the right hand side. This allows to overcome the issue of persistency of excitation for PCE coefficients by resorting to realization trajectory data instead.

Lemma 4. (Stochastic fundamental lemma).
Consider system (2) and its $\mathcal{L}^2(\Omega, \mathcal{F}_k, \mu; \mathbb{R}^{n_z})$, $n_z \in \{n_x, n_u, n_x, n_y\}$ trajectory tuples of random variables, PCE coefficients (5), and the corresponding realizations (1), which are $(\mathbf{X}, \mathbf{U}, \mathbf{W}, \mathbf{Y})_{[0,T-1]}$, $(\mathsf{x},\mathsf{u},\mathsf{w},\mathsf{y})^j_{[0,T-1]}$, $j \in \{0, \dots, L-1\}$, and $(\mathbf{x}, \mathbf{u}, \mathbf{w}, \mathbf{y})_{[0,T-1]}$. Let (1) be controllable with respect to (u, w), then following assertions hold:

(i) Let $(\mathbf{u}, \mathbf{w})_{[0,T-1]}$ be persistently exciting of order $n_x + t$. Then $(\tilde{\mathbf{X}}, \tilde{\mathbf{U}}, \tilde{\mathbf{W}}, \tilde{\mathbf{Y}})_{[0,t-1]}$ satisfies the dynamics (2) if and only if there exists $G \in \mathcal{L}^2(\Omega, \mathcal{F}, \mu; \mathbb{R}^{T-t+1})$ such that

$$\mathcal{H}_t(\mathbf{z}_{[0,T-1]})G = \tilde{\mathbf{Z}}_{[0,t-1]} \qquad (7a)$$

for all $(\mathbf{z}, \tilde{\mathbf{Z}}) \in \{(\mathbf{x}, \tilde{\mathbf{X}}), (\mathbf{u}, \tilde{\mathbf{U}}), (\mathbf{w}, \tilde{\mathbf{W}})\}$.

(ii) Let $(\mathbf{U}, \mathbf{W})_{[0,T-1]}$ satisfy

$$\mathbf{Z}_{[0,T-1]} = \sum_{j=0}^{L-1} \mathsf{z}^j \phi^j, \quad \mathbf{Z} \in \{\mathbf{U}, \mathbf{W}\}$$

with $L \in \mathbb{N}$ and all PCE trajectories $(\mathsf{u}, \mathsf{w})^j_{[0,T-1]}$ with $j \in \{0, \dots, L-1\}$ are persistently exciting of order $n_x + t$. If there exists a $g \in \mathbb{R}^{T-t+1}$ such that, for all $\mathbf{Z} \in \{\mathbf{X}, \mathbf{U}, \mathbf{W}, \mathbf{Y}\}$,

$$\mathcal{H}_t(\mathbf{Z}_{[0,T-1]})g = \tilde{\mathbf{Z}}_{[0,t-1]} \qquad (7b)$$

then $(\tilde{\mathbf{X}}, \tilde{\mathbf{U}}, \tilde{\mathbf{W}}, \tilde{\mathbf{Y}})_{[0,t-1]}$ satisfies the dynamics (2). $\square$

In slightly modified form, the proof of this result is contained in Pan et al. (2021), where the result appears as Lemma 4.

3.3 Discussion

Recall the usual form of a fundamental lemma, i.e., the columns of the Hankel matrix constructed by past variables span the full, finite-horizon system behavior of the LTI system. It is crucial to note that if the Hankel matrix is constructed directly from random variables, this usual form of the lemma does not necessarily hold in the stochastic setting, cf. Part (ii) of Lemma 4. Specifically, notice that upon applying Galerkin projection for $L \in \mathbb{N}$ PCE basis functions to (7a) and using the equivalence of $\operatorname{colsp}\big(\mathcal{H}_t(\mathsf{z}^j_{[0,T-1]})\big)$ and $\operatorname{colsp}\big(\mathcal{H}_t(\mathbf{z}_{[0,T-1]})\big)$ established in (6a), we obtain the system of linear equations to compute the vector $\mathsf{g}^{[0,L-1]}$. This linear problem reads $\big(I_L \otimes \mathcal{H}_t(\mathbf{z}_{[0,T-1]})\big)\,\mathsf{g}^{[0,L-1]} = \tilde{\mathsf{z}}^{[0,L-1]}$, where $\otimes$ denotes the Kronecker product and $\mathsf{g}^{[0,L-1]} \in \mathbb{R}^{L(T-t+1)}$ stacks the vectors g^j into one vector. In contrast, Galerkin projection of (7b) combined with column-space equivalence gives $\big(\mathbf{1}_L \otimes \mathcal{H}_t(\mathbf{z}_{[0,T-1]})\big)\,\mathsf{g} = \tilde{\mathsf{z}}^{[0,L-1]}$,where $\mathbf{1}_L$ is the $L \times 1$ vector of all 1 and g is of dimension $T-t+1$. In other words, Galerkin projection of (7b) does not give sufficient flexibility to represent all trajectories.

Example 1. (If and iff statements in Lemma 4).
We present a simple example illustrating why Part (ii) of Lemma 4 does not admit an *iff* statement. Consider the scalar stochastic system $X_{k+1} = 2X_k + U_k$ with past trajectories given in terms of their PCEs

$$\begin{aligned} X_0 &= 0\phi^0 + 0\phi^1, \quad U_0 = 0\phi^0 + 1\phi^1, \\ X_1 &= 0\phi^0 + 1\phi^1, \quad U_1 = 1\phi^0 + 0\phi^1, \\ X_2 &= 1\phi^0 + 2\phi^1, \quad U_2 = 1\phi^0 + 1\phi^1. \end{aligned}$$

Note that the PCE coefficients of $\mathbf{U}_{[0,2]}$ satisfy the persistency of excitation required by Part (ii) of Lemma 4.

We aim to find $g \in \mathbb{R}^3$ in (7b) to represent $\tilde{X}_0 = 0\phi^0 + 1\phi^1$, $\tilde{U}_0 = 0\phi^0 + 1\phi^1$. We obtain (7b) as

$$\begin{bmatrix} X_0 & X_1 & X_2 \\ U_0 & U_1 & U_2 \end{bmatrix} g = \begin{bmatrix} \tilde{X}_0 \\ \tilde{U}_0 \end{bmatrix}. \tag{8}$$

After Galerkin projection onto the basis functions and stacking the projected equations we obtain $Mg = c$ with

$$M = \left[\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 1 & 1 \\ \hline 0 & 1 & 2 \\ 1 & 0 & 1 \end{array}\right] \quad c = \left[\begin{array}{c} 0 \\ 0 \\ \hline 1 \\ 1 \end{array}\right]$$

where the upper block corresponds to ϕ^0 and the lower one to ϕ^1. By the Rouché–Capelli theorem, $Mg = c$ admits a solution g if and only if $[M|c]$ has the same rank as M. Observe that $\operatorname{rank} M = 3$ and $\operatorname{rank}[M|c] = 4$. Thus, we conclude that (8) does not admit solutions $g \in \mathbb{R}^3$. □

Remark 2. (Knowlegde of past noise realizations).
From an applications point of view further comments are in order. First, observe that Lemma 4 requires knowledge of past realization of noise trajectories. Depending on the context such realization data might be accessible a posteriori through measurements—e.g., consider cases in which the disturbance W_k models volatile renewable energy production or randomly varying energy demands, cf. Bilgic et al. (2022)—or one may estimate them from available measurements. For details on the latter for the case of state feedback and $E = I$ see (Pan et al., 2021). □

Remark 3. (Data-driven stochastic control). One may wonder how to put the presented stochastic fundamental lemma to use for stochastic predictive control leveraging the lemma including numerical examples are provided by Pan et al. (2021). For tailored numerical solutions methods see (Ou et al., 2023), stability results for state feedback can be given (Pan et al., 2022). □

Remark 4. ($\mathcal{L}^2$ Equivalence). From a conceptual point of view, we remark that the PCE approach gives finite-dimensional representations of random variables only in an $\mathcal{L}^2$ equivalence sense. Hence, also the non-parametric system description via the fundamental lemma is in general to be understood in an $\mathcal{L}^2$ equivalence sense. □

4. CONCLUSIONS

This note has recalled first steps towards a fundamental lemma for stochastic LTI systems. To this end, we have recalled parts of the results of our recently submitted paper (Pan et al., 2021). The crucial insights obtained include the column-space equivalence of Hankel matrices for PCE coefficients and realizations which in turn enables the derivation of a fundamental lemma wherein the Hankel matrix contains deterministic data (of realizations or PCE coefficients) while the vector expressing linear combinations of the columns becomes a vector-valued random variable. Moreover, we have shown that the counterpart, i.e. Hankel matrices in terms of random variables and a deterministic vector expressing linear combinations, does not lead to an equivalent description.

REFERENCES

Allibhoy, A. and Cortés, J. (2020). Data-based receding horizon control of linear network systems. *IEEE Control Systems Letters*, 5(4), 1207–1212.

Alsalti, M., Berberich, J., Lopez, V.G., Allgöwer, F., and Müller, M.A. (2021). Data-based system analysis and control of flat nonlinear systems. *arXiv preprint arXiv:2103.02892*.

Berberich, J., Koch, A., Scherer, C.W., and Allgöwer, F. (2020). Robust data-driven state-feedback design. In *2020 American Control Conference (ACC)*, 1532–1538. IEEE.

Bilgic, D., Pan, G., Koch, A., and Faulwasser, T. (2022). Toward data-enabled predictive control of multi-energy distribution systems. *Electric Power Systems Research*. doi: 10.1016/j.epsr.2022.108311. Presented at Power Systems Computation Conference (PSCC). To appear.

Coulson, J., Lygeros, J., and Dörfler, F. (2021). Distributionally robust chance constrained data-enabled predictive control. *IEEE Transactions on Automatic Control*, 6(7), 3289 – 3304.

Coulson, J., Lygeros, J., and Dörfler, F. (2019). Data-enabled predictive control: In the shallows of the DeePC. In *2019 18th European Control Conference (ECC)*, 307–312. IEEE.

De Persis, C. and Tesi, P. (2019). Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, 65(3), 909–924.

De Persis, C. and Tesi, P. (2021). Low-complexity learning of linear quadratic regulators from noisy data. *Automatica*, 128. 109548.

Fristedt, B.E. and Gray, L.F. (2013). *A Modern Approach to Probability Theory*. Springer Science & Business Media.

Lian, Y., Wang, R., and Jones, C.N. (2021). Koopman based data-driven predictive control. *arXiv preprint arXiv:2102.05122*.

Markovsky, I. and Dörfler, F. (2021). Behavioral systems theory in data-driven analysis, signal processing, and control. *Annual Reviews in Control*.

Mühlpfordt, T., Findeisen, R., Hagenmeyer, V., and Faulwasser, T. (2018). Comments on quantifying truncation errors for polynomial chaos expansions. *IEEE Control Systems Letters*, 2(1), 169–174. doi:10.1109/LCSYS.2017.2778138.

Nortmann, B. and Mylvaganam, T. (2020). Data-driven control of linear time-varying systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 3939–3944. IEEE.

Ou, R., Pan, G., and Faulwasser, T. (2023). Data-driven multiple shooting for stochastic optimal control. *IEEE Control Systems Letters*, 7, 313–318. doi:10.1109/LCSYS.2022.3185841.

Pan, G., Ou, R., and Faulwasser, T. (2021). On a stochastic fundamental lemma and its use for data-driven MPC. *Submitted to IEEE TAC, arXiv:2111.13636*.

Pan, G., Ou, R., and Faulwasser, T. (2022). Towards data-driven stochastic predictive control. *International Journal of Robust and Nonlinear Control*. Submitted.

Schmitz, P., Faulwasser, T., and Worthmann, K. (2022). Willems' fundamental lemma for linear descriptor systemsand its use for data-driven output-feedback predictive control. *IEEE Letters of the Control Systems Society*. doi:10.1109/LCSYS.2022.3161054.

Sullivan, T.J. (2015). *Introduction to Uncertainty Quantification*, volume 63. Springer.

van Waarde, H.J., Camlibel, M.K., and Mesbahi, M. (2020a). From noisy data to feedback controllers: Non-conservative design via a matrix S-lemma. *IEEE Transactions on Automatic Control*.

van Waarde, H.J., Eising, J., Trentelman, H.L., and Camlibel, M.K. (2020b). Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11), 4753–4768.

Verhoek, C., Abbas, H.S., Tóth, R., and Haesaert, S. (2021). Data-driven predictive control for linear parameter-varying systems. *IFAC-PapersOnLine*, 54(8), 101–108. 4th IFAC Workshop on Linear Parameter Varying Systems LPVS 2021.

Wiener, N. (1938). The homogeneous chaos. *American Journal of Mathematics*, 897–936.

Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B.L.M. (2005). A note on persistency of excitation. *Systems & Control Letters*, 54(4), 325–329.

Yin, M., Iannelli, A., and Smith, R.S. (2020). Maximum likelihood estimation in data-driven modeling and control. *arXiv preprint arXiv:2011.00925*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Real hyperplane sections and linear series on algebraic curves

Daniel Plaumann *

* *Fakultät für Mathematik, Technische Universität Dortmund, D-44227 Dortmund, Germany*

Abstract: Given a real algebraic curve in projective space, we study the computational problem of deciding whether there exists a hyperplane meeting the curve in real points only. More generally, given any divisor on such a curve, we may ask whether the corresponding linear series contains an effective divisor with totally real support. This translates into a particular type of parametrized real root counting problem that we wish to solve exactly.
We will focus on examples and some general results and conjectures, based on recent work with Huu Phuoc Le and Dimitri Manevich.

Keywords: real algebraic curve, totally real hyperplane section, divisor, Hermite matrix, parametrized root counting

1. INTRODUCTION

In this talk I will survey some computational problems concerning real algebraic curves. Results will be based on a joint paper

Computing totally real hyperplane sections and linear series on algebraic curves
by Huu Phuoc Le, Dimitri Manevich, Daniel Plaumann

available from

`https://arxiv.org/abs/2106.13990`

and accepted for publication in Le Matematiche.

From the introduction:

Given a real algebraic curve X of degree d embedded into some projective space, we consider the computational problem of deciding whether there exists a real hyperplane meeting X in a prescribed number r of real points, counted with multiplicity. Of particular interest is the case $r = d$, i.e., hyperplanes meeting X in real points only. More generally, given any divisor D on X defined over $\mathbb{R}$, and thus consisting of real points and complex-conjugate pairs, we may ask whether the linear series $|D|$ contains an effective divisor with totally real support. (The first question is the special case when D is a hyperplane section of a suitably embedded curve.) A number of general results have been obtained in this direction: The answer is known to be positive for any divisor of sufficiently high degree (see Scheiderer (2000)). However, the precise degree required, relative to the genus of X, is the subject of several results and conjectures, some of which we will investigate from a computational point of view. Explicit bounds are only known if the real locus $X(\mathbb{R})$ has many connected components (so-called M-curves or $(M-1)$-curves), by results due to Huisman (2001) and Monnier (2003). On the other hand, very little is known about curves whose number of connected components is not close to maximal.

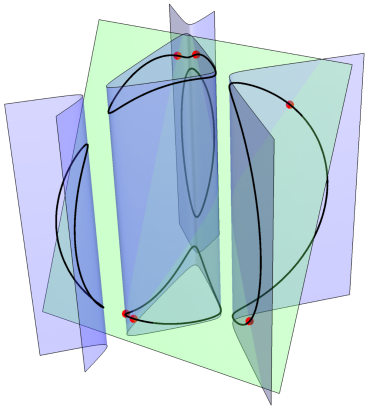


Fig. 1. A real space curve of degree 6 with a totally real hyperplane section.

Of course, the computational problem makes sense for any given curve and divisor, regardless of whether or not there is a general result covering all curves and divisors of the given kind.

It comes down to "solving" polynomial systems whose coefficients depend on parameters. More precisely, we consider the coefficients of the equation defining the hyperplane as *parameters*. One then associates a hyperplane to a point in the space of parameters. The number of real points at the intersection of the considered hyperplane with the curve may vary depending on the parameters, while the number of complex intersection points between the curve and the hyperplane is equal to the degree d for *generic* values of the parameters. (If the points are counted with intersection multiplicities and the curve is not contained in a hyperplane, this complex intersection number is equal to d for *all* values of the parameters.) Hence, from a computational point of view, we are considering a polynomial system, depending on parameters such that, when these

parameters take generic values, the solution set over the complex numbers is finite. When the input system generates a radical ideal, the algorithm we use , which is detailed in Le and Safey El Din (2020), computes a partition of a dense semi-algebraic subset of the space of parameters into open semi-algebraic sets such that the number of real *simple* solutions (i.e., without multiplicities) to the input system is invariant for any point chosen in one of these sets. To do this, we compute a symmetric matrix called the *parametric Hermite matrix*, whose entries are polynomials depending on the parameters and such that, after specialization, its signature coincides with the number of real solutions to the specialized system. This allows us to classify the possible number of real roots to the input system with respect to the parameters.

Our main findings can be summarized as follows.

1. There exist canonical curves X in $\mathbb{P}^3$ with one or two ovals which do not allow simple totally real hyperplane sections.
2. There exists a curve X in $\mathbb{P}^3$ of genus two and degree five having one oval which does not allow a simple totally real hyperplane section.
3. There are infinitely many plane quartics X with many ovals possessing a (complete) linear series of degree four which does not contain a totally real divisor.
4. For every $d \geq 3$ and every number $1 \leq s \leq g+1$ with $g = \frac{(d-1)(d-2)}{2}$, there exists a plane curve X of degree d, genus g and having s branches such that the linear series of lines $|L|$ is totally real.

REFERENCES

J. Huisman. On the geometry of algebraic curves having many real components. *Rev. Mat. Complut.*, 14(1):83–92, 2001.

J. Huisman. Non-special divisors on real algebraic curves and embeddings into real projective spaces. *Ann. Mat. Pura Appl. (4)*, 182(1):21–35, 2003.

A. Kobel, F. Rouillier, and M. Sagraloff. Computing real roots of real polynomials ... and now for real! In *Proceedings of the ACM on International Symposium on Symbolic and Algebraic Computation*, ISSAC '16, page 303–310, New York, NY, USA, 2016. Association for Computing Machinery.

V. A. Krasnov. Albanese mapping for $GM\mathbf{Z}$-varieties. *Mat. Zametki*, 35(5):739–747, 1984.

M. Kummer and D. Manevich. On Huisman's conjectures about unramified real curves. *Preprint arXiv:1909.09601*, 2019.

H. P. Le and M. Safey El Din. Solving parametric systems of polynomial equations over the reals through hermite matrices. *Preprint arXiv:2011.14136*, 2020.

J.-P. Monnier. Divisors on real curves. *Adv. Geom.*, 3(3):339–360, 2003.

J.-P. Monnier. On real generalized Jacobian varieties. *J. Pure Appl. Algebra*, 203(1-3):252–274, 2005.

M. Safey El Din and E. Schost. Polar varieties and computation of one point in each connected component of a smooth real algebraic set. In *Proc. of the 2003 Int. Symp. on Symb. and Alg. Comp.*, ISSAC '03, page 224–231, NY, USA, 2003. ACM.

C. Scheiderer. Sums of squares of regular functions on real algebraic varieties. *Trans. Amer. Math. Soc.*, 352(3):1039–1069, 2000.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On d-Collision-Free Dynamical Systems ⋆

Melanie Harms * **Simone Bamberger** ** **Eva Zerz** ***
Michael Herty ****

* *Lehrstuhl für Algebra und Zahlentheorie, RWTH Aachen University, Aachen, Germany (e-mail: melanie.harms@rwth-aachen.de)*
** *Lehrstuhl für Algebra und Zahlentheorie, RWTH Aachen University, Aachen, Germany (e-mail: simone.bamberger@rwth-aachen.de)*
*** *Lehrstuhl für Algebra und Zahlentheorie, RWTH Aachen University, Aachen, Germany (e-mail: eva.zerz@math.rwth-aachen.de)*
**** *Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, Aachen, Germany (e-mail: herty@igpm.rwth-aachen.de)*

Abstract: This extended abstract presents several recent results and generalizations that have been obtained in the theory of collision-freeness studied in Zerz and Herty (2019). A nonlinear ODE system $\dot{x}(t) = f(x(t))$ is called collision-free if the solution to the initial value problem with $x(0) = x^0$ has distinct components for all times t whenever the initial state x^0 has distinct components. This is an important structural property of particle systems. Here, we address the case where the state of the i-th particle has d components x_{ik}, and a collision occurs if there exists t and $i \neq j$ such that $x_{ik}(t) = x_{jk}(t)$ for all $1 \leq k \leq d$.

Keywords: Structural properties; multivariable systems; linear/polynomial/nonlinear systems.
AMS Codes: 93B25, 93B27.

1. PRELIMINARIES AND KNOWN RESULTS

Let $N \geq 1$ be an integer, $U \subseteq \mathbb{R}^N$ an open set, and let $f : U \to \mathbb{R}^N$ be a $\mathcal{C}^1$-function. Consider the ordinary differential equation (ODE) $\dot{x}(t) = f(x(t))$. Let $x^0 \in U$ be given. The initial value problem (IVP) $\dot{x} = f(x)$, $x(0) = x^0$ has a unique nonextendable $\mathcal{C}^1$-solution

$$\varphi(\cdot, x^0) : J(x^0) \to U,$$

where $J(x^0) \subseteq \mathbb{R}$ is an open interval containing $t_0 = 0$, called the maximal existence interval of the IVP. A subset $S \subseteq U$ is called invariant for $\dot{x} = f(x)$ if $x^0 \in S$ implies that $\varphi(t, x^0) \in S$ for all $t \in J(x^0)$. Clearly, a set S is invariant if and only if its complement $U \setminus S$ is invariant.

We are particularly interested in the case where the set S is given by polynomial equations. Given a subset $P \subseteq \mathcal{P} := \mathbb{R}[X_1, \ldots, X_N]$, we write

$$V := \mathcal{V}(P) = \{x \in \mathbb{R}^N \mid p(x) = 0 \text{ for all } p \in P\}$$

for the variety defined by P. Conversely, the set

$$\mathcal{J}(V) = \{p \in \mathcal{P} \mid p(x) = 0 \text{ for all } x \in V\}$$

is an ideal in $\mathcal{P}$. We have $\mathcal{J}(V_1 \cup V_2) = \mathcal{J}(V_1) \cap \mathcal{J}(V_2)$ and $\mathcal{V}(I_1 \cap I_2) = \mathcal{V}(I_1) \cup \mathcal{V}(I_2)$ for varieties $V_i \subseteq \mathbb{R}^N$ and ideals $I_i \subseteq \mathcal{P}$. Since $\mathcal{P}$ is Noetherian, any ideal in $\mathcal{P}$ is finitely generated.

The Lie derivative of $p \in \mathcal{P}$ along f is defined by

$$L_f(p) = \sum_{i=1}^{N} \frac{\partial p}{\partial X_i} f_i \in \mathcal{C}^1(U, \mathbb{R}).$$

⋆ This work was supported by DFG-SFB 1481 and DFG-TRR 195.

In the special case where f is a polynomial function, which will be identified with some $f \in \mathcal{P}^N$, we set $U = \mathbb{R}^N$ and we have $L_f(p) \in \mathcal{P}$ for all $p \in \mathcal{P}$.

The following two results are folklore, but we state them for the sake of self-containedness; see also Zerz and Walcher (2012) and Harms et al. (2017).

Lemma 1. Let $f \in \mathcal{P}^N$ and a variety $V \subseteq \mathbb{R}^N$ be given. Then V is invariant for $\dot{x} = f(x)$ if and only if $L_f(\mathcal{J}(V)) \subseteq \mathcal{J}(V)$.

The following basic facts from algebraic geometry can be found in standard textbooks such as Kunz (1985). A variety V is said to be irreducible if it cannot be written as a union $V = V_1 \cup V_2$ with subvarieties $V_i \subsetneq V$. A maximal irreducible subvariety of V is called an irreducible component of V. Any variety V has only finitely many irreducible components V_i, and V can be written as a union of its irreducible components such that none of the V_i is superfluous. A variety V is irreducible if and only if $\mathcal{J}(V)$ is a prime ideal. Recall that an ideal $\mathfrak{p} \subsetneq \mathcal{P}$ is called prime if $pq \in \mathfrak{p}$ implies $p \in \mathfrak{p}$ or $q \in \mathfrak{p}$.

Theorem 2. Let $f \in \mathcal{P}^N$ and a variety $V \subseteq \mathbb{R}^N$ be given. Let $V = \bigcup V_i$ be the decomposition of V into its irreducible components. Then V is invariant for $\dot{x} = f(x)$ if and only if each V_i is.

With these preparations, we now introduce the central concept of this contribution.

Definition. Let $n \geq 2$ and $d \geq 1$ be integers. A vector $x = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{nd}$ with $x_i = (x_{i1}, \ldots, x_{id})^T$ is said to have *d-distinct components* if the x_i are (pairwise) distinct. An ODE $\dot{x} = f(x)$ is called *d-collision-free* if for any vector x^0 with d-distinct components, the solution

$\varphi(t, x^0)$ to the IVP $\dot{x} = f(x)$, $x(0) = x^0$ has d-distinct components for all t in the maximal existence interval. A matrix $A \in \mathbb{R}^{nd\times nd}$ is called *d-collision-free* if the linear ODE $\dot{x} = Ax$ is d-collision-free.

Let $I_{ij} := \langle X_{i1} - X_{j1}, \dots, X_{id} - X_{jd}\rangle \subseteq \mathcal{P}$, where

$$\mathcal{P} = \mathbb{R}[X_{11}, \dots, X_{1d}, \dots, X_{n1}, \dots, X_{nd}]$$

is a polynomial ring in nd variables, and let

$$V_{ij} := \mathcal{V}(I_{ij}) = \{x \in \mathbb{R}^{nd} \mid p(x) = 0 \text{ for all } p \in I_{ij}\} \subset \mathbb{R}^{nd}$$

be its vanishing set. Finally, let

$$V = \bigcup_{1\le i<j\le n} V_{ij}. \tag{1}$$

Clearly $\complement V := \mathbb{R}^{nd} \setminus V$ is the set of all vectors with d-distinct components. By definition, $\dot{x} = f(x)$ is d-collision-free if and only if $U \setminus V$ is an invariant set of this ODE. Equivalently, $U \cap V$ itself is an invariant set of $\dot{x} = f(x)$.

Since (1) is the decomposition of V into irreducible components, we may conclude from Theorem 2 that $\dot{x} = f(x)$ with $f \in \mathcal{P}^{nd}$ is d-collision-free if and only if each V_{ij} is invariant, which means that $x_i^0 = x_j^0$ implies that $x_i(t) = x_j(t)$ for all t in the maximal existence interval of the solution x with $x(0) = x^0$.

2. LINEAR AND POLYNOMIAL SYSTEMS

In this section, we first analyze the d-collision-freeness of linear systems, that is, $\dot{x} = Ax$ with $A \in \mathbb{R}^{nd\times nd}$.

Notation. Let $\underline{n}$ denote the set $\{1, \dots, n\}$. For a matrix $A \in \mathbb{R}^{nd\times nd}$, we consider the partition of A into submatrices $A_{ij} \in \mathbb{R}^{d\times d}$ for $i, j \in \underline{n}$.

Theorem 3. Let $A \in \mathbb{R}^{nd\times nd}$ be given. The following are equivalent:

(i) A is d-collision-free.
(ii) The submatrices A_{ij} of A satisfy

$$A_{ik} = A_{jk} \quad \text{for all } i, j, k \in \underline{n} \text{ with } i \neq k, j \neq k \tag{2}$$

and

$$\sum_{k=1}^{n} A_{ik} = \sum_{k=1}^{n} A_{jk} \quad \text{for all } i, j \in \underline{n}. \tag{3}$$

Note that in view of (2), Equation (3) is equivalent to

$$A_{ii} + A_{ij} = A_{ji} + A_{jj} \quad \text{for all } i, j \in \underline{n}.$$

From the theorem, we derive the following insights into the structural properties of d-collision-free matrices.

Corollary 4. Let $R \subseteq \mathbb{R}^{nd\times nd}$ denote the set of all d-collision-free matrices. Then R is a subring of $\mathbb{R}^{nd\times nd}$. Any $A \in R$ is uniquely determined by $A_{11}, \dots, A_{1n}$ and A_{21}, where these matrices can be freely chosen. Thus the dimension of R as a real vector space equals $(n + 1)d^2$.

Corollary 5. Let $f \in \mathcal{P}^{nd}$ be given. Then $\dot{x} = f(x)$ is d-collision-free if and only if each $f_{ik} - f_{jk}$ is contained in I_{ij}, where $1 \le i < j \le n$ and $1 \le k \le d$.

3. GENERAL $\mathcal{C}^1$-SYSTEMS

Next, we consider the general $\mathcal{C}^1$-case.

Consider $\dot{x} = f(x)$, where $f \in \mathcal{C}^1(U, \mathbb{R}^{nd})$ for some open set $U \subseteq \mathbb{R}^{nd}$. Let $V_{ij} = \bigcap_{k=1}^{d} \mathcal{V}(X_{ik} - X_{jk})$ and $V = \bigcup_{1\le i<j\le n} V_{ij}$. Recall that $\complement V = \mathbb{R}^{nd} \setminus V$ is the set of all vectors with d-distinct components. This is an open and dense subset of $\mathbb{R}^{nd}$ and its boundary equals V.

Lemma 6. Let $f \in \mathcal{C}^1(U, \mathbb{R})$ for some open set $U \subseteq \mathbb{R}^{nd}$. Suppose that f vanishes on $U \cap \mathcal{V}(X_{11}, \dots, X_{1d})$. Then there exist functions $a_i \in \mathcal{C}^0(U, \mathbb{R})$ such that $f(x) = a_1(x)x_{11} + \ldots + a_d(x)x_{1d}$ for all $x = (x_1^T, \dots, x_n^T)^T \in U$ and $x_1 = (x_{11}, \dots, x_{1d})^T$.

Lemma 7. (a) Let $q_1, \dots, q_l \in \mathcal{P}$ be polynomials. If $U \cap \mathcal{V}(q_1, \dots, q_l)$ is an invariant set of $\dot{x} = f(x)$, then every $L_f(q_k)$, where $1 \le k \le l$, vanishes on $U \cap \mathcal{V}(q_1, \dots, q_l)$.

(b) $U \cap V_{ij}$ is an invariant set of $\dot{x} = f(x)$ if and only if every $f_{ik} - f_{jk}$, where $1 \le k \le d$, vanishes on $U \cap V_{ij}$.

Theorem 8. $U \cap V$ is an invariant set if and only if each $U \cap V_{ij}$ is an invariant set for $1 \le i < j \le n$.

Corollary 9. The ODE $\dot{x} = f(x)$, where $f \in \mathcal{C}^1(U, \mathbb{R}^{nd})$ for some open set $U \subseteq \mathbb{R}^{nd}$, is d-collision-free if and only if each $f_{ik} - f_{jk}$ vanishes on $U \cap V_{ij}$ for $1 \le i < j \le n$ and $1 \le k \le d$.

Example. Consider

$$\dot{x}_i = \sum_{k=1}^{n} \phi(x_k - x_i) \tag{4}$$

for some $\phi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$ and $1 \le i \le n$. This system is d-collision-free.

4. DEPENDENCE ON d AND THE ROLE OF PERMUTATION SYMMETRY

Corollary 10. Let $N = nd = n'd'$, where d is a multiple of d'. Then d'-collision-freeness of an N-dimensional system $\dot{x} = f(x)$ implies d-collision-freeness. In particular, 1-collision-freeness (as studied in Zerz and Herty (2019)) implies d-collision-freeness for all divisors d of the state space dimension N.

Let $G = \{P \otimes I_d \mid P \text{ is an } n \times n \text{ permutation matrix}\} \subseteq \mathbb{R}^{nd\times nd}$, where I_d is the $d\times d$ identity matrix and $\otimes$ denotes the Kronecker product.

Definition. A function $f \in \mathcal{C}^1(U, \mathbb{R}^{nd})$ is called *G-symmetric* if $f(Tx) = Tf(x)$ holds for all $T \in G$ and $x \in U$ with $Tx \in U$. Writing $f = (f_1, \dots, f_n)$ with $f_i \in \mathcal{C}^1(U, \mathbb{R}^d)$, this means that

$$f_{\pi(i)}(x_1, \dots, x_n) = f_i(x_{\pi(1)}, \dots, x_{\pi(n)})$$

holds for any permutation $\pi \in S_n$ of the n substates and any $x = (x_1^T, \dots, x_n^T)^T \in U$ such that both sides of the equation are well-defined.

It turns out that many common models of particle interactions, such as (4), are permutation symmetric in this sense. This reflects the assumption that the particles are identical and indiscernible.

Example. Suppose that $U = V^n$ for some open set $V \subseteq \mathbb{R}^d$. Consider

$$f_i(x) = f_i(x_1, \dots, x_n) = \chi(x_i) + \sum_{k=1, k\neq i}^{n} \psi(x_k, x_i)$$

for some $\chi \in \mathcal{C}^1(V, \mathbb{R}^d)$, $\psi \in \mathcal{C}^1(V \times V, \mathbb{R}^d)$ and $1 \leq i \leq n$. Then f is G-symmetric.

Theorem 11. If f is G-symmetric, then $\dot{x} = f(x)$ is d-collision-free.

Corollary 12. Let $A \in \mathbb{R}^{nd\times nd}$ be given and let $A_{ij} \in \mathbb{R}^{d\times d}$ for $i, j \in \underline{n}$ denote its submatrices as in Theorem 3. Then A is G-symmetric, that is, it commutes with every $T \in G$, if and only if A is both d-collision-free and block symmetric, that is, $A_{ij} = A_{ji}$ for all $i, j \in \underline{n}$.

Both conditions are equivalent to

$$A = \begin{bmatrix} C & B & \dots & B \\ B & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & B \\ B & \dots & B & C \end{bmatrix}$$

for some $B, C \in \mathbb{R}^{d\times d}$.

5. OUTLOOK AND FUTURE WORK

We expect the theory of d-collision-freeness to deepen the understanding of ODE particle systems which arise with traffic, swarm, or consensus models, for instance, see Albi et al. (2015) or Miao et al. (2016). In particular, we are interested in studying how this property carries over to the PDE model resulting in the mean-field kinetic limit considered e.g. in Herty et al. (2015); Tordeux et al. (2018).

REFERENCES

Albi, G., Herty, M., and Pareschi, L. (2015). Kinetic description of optimal control problems and applications to opinion consensus. *Commun. Math. Sci.*, 13, 1407–1429.

Bamberger, S. (2021). d-Collision-Free Dynamical Systems. Master's Thesis, RWTH Aachen University.

Harms, M., Schilli, C., and Zerz, E. (2017). Polynomial control systems: invariant sets given by algebraic equations/inequations. *IFAC-PapersOnLine*, 50, 677–680.

Herty, M., Pareschi, L., and Steffensen, S. (2015). Mean-field control and Riccati equations. *Netw. Heterog. Media*, 10, 699–715.

Kunz, E. (1985). *Introduction to Commutative Algebra and Algebraic Geometry*. Birkhäuser.

Miao, Z., Wang, Y., and Fierro, R. (2016). Collision-free consensus in multi-agent networks: a monotone systems perspective. *Automatica*, 64, 217–225.

Tordeux, A., Costeseque, G., Herty, M., and Seyfried, A. (2018). From traffic and pedestrian follow-the-leader models with reaction time to first order convection-diffusion flow models. *SIAM J. Appl. Math.*, 78, 63–79.

Zerz, E. and Herty, M. (2019). Collision-Free Dynamical Systems. *IFAC-PapersOnLine*, 52, 72–76.

Zerz, E. and Walcher, S. (2012). Controlled invariant hypersurfaces of polynomial control systems. *Qual. Theory Dyn. Syst.*, 11, 145–158.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Port-Hamiltonian modeling of interacting particle systems

Birgit Jacob * **Claudia Totzeck** **

* *Bergische Universität Wuppertal, Fakultät für Mathematik und Naturwissenschaften, IMACM, Arbeitsgruppe Funktionalanalysis, Gaußstraße 20, D-42119 Wuppertal, Germany (e-Mail: bjacob@uni-wuppertal.de).*
** *Bergische Universität Wuppertal, Fakultät für Mathematik und Naturwissenschaften, IMACM, Arbeitsgruppe Optimierung, Gaußstraße 20, D-42119 Wuppertal, Germany (e-Mail: totzeck@uni-wuppertal.de).*

Abstract: A port-Hamiltonian formulation of a general class of interacting particle systems and its corresponding mean-field partial-differential equation is discussed. To establish the port-Hamiltonian structure of the interacting particle systems a specific variable transformation is employed. It turns out that an appropriate retransformation of the characteristics corresponding to the mean-field partial differential equation yields again a port-Hamiltonian structure.

Keywords: Port-Hamiltonian modeling, port-Hamiltonian distributed parameter systems, interacting particle systems, mean-field limit, multi-agent systems.

1. INTRODUCTION

Port-based network modeling of complex physical systems leads to port-Hamiltonian systems (PHS). For finite-dimensional systems there is by now a well-established theory, see van der Schaft (2006); Eberard et al. (2007); Duindam et al. (2009). The port-Hamiltonian approach has been further extended to the infinite-dimensional situation in van der Schaft and Maschke (2002); Zwart et al. (2010); Villegas (2007); Jacob and Zwart (2012). This class is further closed under network interconnection. That is, coupling of port-Hamiltonian systems again leads to a port-Hamiltonian system. Furthermore, the port-Hamiltonian approach is suitable for the investigation of the qualitative solution behavior and optimization questions, as it provides an energy balance.

In this talk we model interacting particle systems as well as the corresponding mean-field partial-differential equation as port-Hamiltonian system. A recent work by Matei et al. (2019) discusses the port-Hamiltonian formulation of a Cucker-Smale dynamic involving three particles. The key idea leading the reformulation is the interpretation of the Cucker-Smale interactions as generalized spring-damper systems. The approach has several advantages: the introduction of relative positions factors out the translational invariance of the system, and, the port-Hamiltonian structure allows to identify several conserved quantities namely the Hamiltonian and the so-called Casimir functions. In future work, the novel interpretation of interacting particle dynamics as generalized spring-damper systems with PHS structure can be employed for new control-strategies of interacting particle systems.

2. MODELING

We recall the classical formulation of interacting particle systems in position and velocity coordinates before we introduce the reformulation in PHS structure.

Let us denote the space dimension d and consider $N \in \mathbb{N}$ agents, let $x_i \colon [0,T] \to \mathbb{R}^d$ and $v_i \colon [0,T] \to \mathbb{R}^d$ denote the position and velocity functions of the i-th agent, respectively. We collect the positions and the velocities of all agents in the vectors x and v with $[x]_i = x_i$ and $[v]_i = v_i$ for $i = 1, \dots, N$, respectively. Let $G \in \mathcal{C}(\mathbb{R}, \mathbb{R}_{\geq 0})$ model a generalized damper and $V \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ be the potential for the binary interactions of the particles. For the forces resulting from the interactions we require

$$\nabla_{x_i} V(x_j - x_i) = -\nabla_{x_j} V(x_j - x_i). \tag{1}$$

According to Newton's second law, a general interacting particle systems can be written as

$$\frac{d}{dt} x = v, \qquad x(0) = x_0, \tag{2a}$$

$$\frac{d}{dt} v = \mathcal{G}(x)v - \nabla \mathcal{V}(x), \qquad v(0) = v_0, \tag{2b}$$

where

$$[\mathcal{G}(x)]_{ii} = -\frac{1}{N} \sum_{j \neq i} G(|x_j - x_i|)$$

$$[\mathcal{G}(x)]_{ij} = \frac{1}{N} G(|x_j - x_i|), \quad i \neq j,$$

model pairwise alignment of the agents, like for example in the model by Cucker and Smale (2007), and

$$[\nabla \mathcal{V}(x)]_{ii} = 0, \quad [\nabla \mathcal{V}(x)]_{ij} = \frac{1}{N} \nabla_{x_i} V(x_j - x_i), \quad i \neq j$$

model pairwise interactions of the particles. We remark, that the matrix $\mathcal{G}(x)$ is negative semi-definite by definition. For well-posedness the systems is supplemented with initial conditions $x(0) = x_0, v(0) = v_0$ with $[x_0]_i$ and $[v_0]_i$ drawn independently from a probability distribution $\hat{f} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$.

This general class of interaction models contains the following well-known examples: Cucker and Smale (2007), Matei et al. (2019), Morse-interactions (sheep flocks, double and single milling birds) as proposed in D'Orsogna et al. (2006), or herding dynamics Burger et al. (2020).

For later use we mention the well-known mean-field PDE corresponding to (2) given by

$$\begin{aligned} \partial_t f_t + \nabla_x(v f_t) &= \nabla_v \left((G * \varrho_t) + \nabla_x V * \varrho_t) f_t \right), \qquad (3) \\ f_0(x,v) &= \hat{f}(x,v) \end{aligned}$$

for some probability measure $\hat{f} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, where we used $\varrho_t(x) = \int_{\mathbb{R}^d} f_t(x,v)dv$.

We note that the actual positions of the agents do not influence the dynamic but rather the relative distances $x_j - x_i$. This is exploited in the derivation of port-Hamiltonian system (PHS) reformulation proposed below. Indeed, following the approach of Matei et al. (2019) we introduce new variables $q_{ij} = x_i - x_j$ and $p_i = mv_i$, the relative positions and impulses of the agents. In the remainder we set $m = 1$ such that $p_i = v_i$ and write $z = (q,p)$ with $[q]_{ij} = q_{ij}$ and $[p]_i = p_i$. For notational convenience we fix the ordering of q as follows:
We consider the matrix Q with $[Q]_{ij} = q_{ij}$ for $i \neq j$ and $[Q]_{ii} = 0$. We "remove" the diagonal of the matrix and write the entries q_{ij} of the resulting matrix $Q_- \in (\mathbb{R}^d)^{N\times(N-1)}$ row-wise in the vector q.

Using the new variables the above dynamic can be written as

$$\frac{d}{dt} q_{ij} = p_i - p_j, \quad q_{ij}(0) = (q_0)_{ij}, \qquad (4a)$$

$$\frac{d}{dt} p = \mathcal{G}(q)p - \nabla\mathcal{V}(q), \quad p(0) = v_0, \qquad (4b)$$

where

$$\begin{aligned} [\mathcal{G}(q)]_{ii} &= -\frac{1}{N}\sum_{j\neq i} G(|q_{ij}|) \\ [\mathcal{G}(x)]_{ij} &= \frac{1}{N} G(|q_{ij}|), \quad i \neq j, \\ [\nabla\mathcal{V}(q)]_{ii} &= 0, \quad [\nabla\mathcal{V}(q)]_{ij} = \frac{1}{N}\nabla V(q_{ij}), \quad i \neq j. \end{aligned}$$

Remark 1. In order to be consistent, the initial conditions of (2) and (4) have to be chosen carefully. In case of independent uniformly distributed initial positions $\text{law}(x_i(0)) = \mathcal{U}[a,b]$ with $a,b \in \mathbb{R}, a < b$, we can exploit the fact that the probability distribution function of $x_i(0) - x_j(0)$ is given by a convolution, leading to $\text{law}(q_{ij}(0)) = \mathcal{U}[a,b] * \mathcal{U}[-b,-a]$.

As the Hamiltonian reformulation for general N is quite technical, we sketch the main ideas with the derivation of the PHS formulation for $N = 2$:

Let $z = (q_{12}, q_{21}, p_1, p_2)^\top$ and consider the Hamiltonian

$$\mathcal{H}(z) = \frac{1}{2}|p_1|^2 + \frac{1}{2}|p_2|^2 + \frac{1}{4}(V(q_{12}) + V(q_{21})).$$

The dynamics of z is defined by (4) leading to

$$\frac{d}{dt} z = \begin{pmatrix} p_1 - p_2 \\ p_2 - p_1 \\ \mathcal{G}(q)p - \nabla\mathcal{V}(q) \end{pmatrix}$$

Note that it holds

$$\frac{\partial\mathcal{H}(z)}{\partial z} = (\frac{1}{4}\nabla V(q_{12}), \frac{1}{4}\nabla V(q_{21}), p_1, p_2)^\top$$

which allows us to rewrite

$$\begin{aligned} \frac{d}{dt} z &= \begin{pmatrix} 0 & \begin{pmatrix} I & -I \\ -I & I \end{pmatrix} \\ \begin{pmatrix} -I & I \\ I & -I \end{pmatrix} & \mathcal{G}(q) \end{pmatrix} \begin{pmatrix} \frac{1}{4}\nabla V(q) \\ p \end{pmatrix} + \begin{pmatrix} 0 \\ B \end{pmatrix} u \\ &= \left[J - \begin{pmatrix} 0 & 0 \\ 0 & -\mathcal{G}(q) \end{pmatrix} \right] \frac{\partial\mathcal{H}(z)}{\partial z} \end{aligned}$$

with

$$J = \begin{pmatrix} 0 & \begin{pmatrix} I & -I \\ -I & I \end{pmatrix} \\ \begin{pmatrix} -I & I \\ I & -I \end{pmatrix} & 0 \end{pmatrix}.$$

Here I denotes the identity matrix in $\mathbb{R}^{d\times d}$. We want to emphasize that the upper-right part of J is set by the structure of q_{ij} and by the skew-symmetry of J this also sets the lower-left part of J. Therefore assumption (1) is crucial for port-Hamiltonian reformulation.

For $N \in \mathbb{N}$ the above derivation generalizes to $z = (q,p)^\top$ and

$$\mathcal{H}(z) = \frac{1}{2}p^T p + \frac{1}{2N}\sum_{i=1}^{N}\sum_{j\neq i}^{N} V(q_{ij}).$$

We then derive

$$\frac{d}{dt} z = \left[\begin{pmatrix} 0 & J \\ -J & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & -\mathcal{G}(q) \end{pmatrix} \right] \frac{\partial\mathcal{H}(z)}{\partial z}$$

where $J \in (\mathbb{R}^{d\times d})^{N(N-1)\times N}$ is given by

$$J = \begin{pmatrix} J_1 \\ \vdots \\ J_N \end{pmatrix}$$

with

$J_i \in \mathbb{R}^{(N-1)\times N}$, $i = 1, \ldots, N$,

$$[J_i]_{jk} = \begin{cases} I, & k = i \\ -I, & (k = j+1 \wedge j \geq i) \vee (k = j < i) \\ 0, & \text{otherwise} \end{cases}.$$

In particular

$$J_1 = \begin{pmatrix} I & -I & 0 & 0 & \cdots & 0 \\ I & 0 & -I & 0 & \cdots & 0 \\ \vdots & & & & & \vdots \\ I & & & & & -I \end{pmatrix}$$

and

$$J_2 = \begin{pmatrix} -I & I & 0 & 0 & \cdots & 0 \\ 0 & I & -I & 0 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & I & & & & -I \end{pmatrix}$$

Then it is easy to verify that (4) can be written as

$$\frac{d}{dt}z = (J - R(z))\frac{\partial \mathcal{H}(z)}{\partial z}, \qquad z(0) = (q(0), p(0)).$$

At first sight we have increased the number of variables from $2Nd$ to dN^2 during the reformulation process. Note that due to the dependencies, it is enough to have the information of one particle and then the relative positions of all the other particles to recover the full information of the system. This observation together with the fact that the PHS formulation factors out the translational invariance of the system, motivates us to introduce the center of mass variable $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and the relative positions $\xi_i = \frac{1}{N}\sum_{j=1, j\neq i}^{N} q_{ij} = x_i - \bar{x}$. System (4) can then by rewritten as

$$\frac{d}{dt}\xi_i = p_i - \frac{1}{N}\sum_{j=1}^{N} p_j, \tag{5}$$

$$\frac{d}{dt}p_i = \frac{1}{N}\sum_{j=1}^{N}\left[\mathcal{G}(\xi_i - \xi_j)p_j - \nabla\mathcal{V}(\xi_i - \xi_j)\right] \tag{6}$$

supplemented with initial conditions $\xi_i(0) = x_i(0) - \bar{x}(0)$ and $p_i(0) = v(0)$.

Remark 2. It is important to note that we loose the port-Hamiltonian structure with this reformulation. Indeed, the right-hand side of (5) sets the structure of the upper-right part of J. Nonlinearities of ∇V prevent a suitable reformulation of (6) to meet this structure. We further remark that in general the nonlinearity of ∇V prevents to reduce the number of variables q_{ij} to only N variables.

3. MEAN-FIELD LIMIT

For interacting particle systems in x-v-formulation a model hierarchy exists Carrillo et al. (2010) that ranges from the particle description, to a mesoscopic and even a hydrodynamic formulation. In the following we discuss the mean-field limit in the port-Hamiltonian formulation. We use the transformed version in (ξ, p) coordinates. It is important to note that $q_{ij} = \xi_i - \xi_j$.

Let $\delta(x, y)$ denote the Delta-distribution on $\mathbb{R}^d \times \mathbb{R}^d$ and define the empirical measure

$$f^N(t, \xi, p) = \frac{1}{N}\sum_{i=1}^{N}\delta(\xi - \xi_i, p - p_i).$$

Further, let $\varphi \in C^1(\mathbb{R}^d \times \mathbb{R}^d)$ be an arbitrary test function. To compute the evolution of f^N, we use the notation

$$\langle f^N, \varphi\rangle = \int_{\mathbb{R}^d\times\mathbb{R}^d}\varphi(\xi, p)df^N(\xi, p) = \frac{1}{N}\sum_{i=1}^{N}\varphi(\xi_i, p_i)$$

and (formally) obtain

$$\begin{aligned}
\left\langle \partial_t f^N, \varphi\right\rangle &= \frac{d}{dt}\left\langle f^N, \varphi\right\rangle \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\nabla_p\varphi(p_i, \xi_i)\cdot\frac{d}{dt}p_i + \nabla_\xi\varphi(p_i, \xi_i)\cdot\frac{d}{dt}\xi_i\right) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\nabla_p\varphi(p_i, \xi_i)\cdot\left[\frac{1}{N}\sum_{j\neq i}G(\xi_i - \xi_j)p_i\right] \\
&\quad - \frac{1}{N}\sum_{i=1}^{N}\nabla_p\varphi(p_i, \xi_i)\cdot\left[\frac{1}{N}\sum_{j\neq i}\nabla V(\xi_i - \xi_j)\right] \\
&\quad + \nabla_\xi\varphi(p_i, \xi_i)\cdot\left[\frac{1}{N}\sum_{j\neq i}p_i - p_j\right] \\
&= \left\langle \nabla_p\cdot\left(\int G(\xi - \bar{\xi})p - \nabla V(\xi - \bar{\xi})\, df_t^N(\bar{p}, \bar{\xi}))\, f_t^N(p, \xi)\right)\right. \\
&\quad \left. - \nabla_\xi\cdot\left(\left[p - \int\bar{p}\, df_t^N(\bar{p}, \bar{\xi})\right] f_t^N(p, \xi)\right), \varphi\right\rangle.
\end{aligned}$$

With the help of the Variational Lemma we can identify the evolution equation for f^N as

$$\begin{aligned}
&\partial_t f_t^N + \nabla_\xi\cdot\left(\left[p - \int\bar{p}\, df_t^N(\bar{p}, \bar{\xi})\right] f_t^N(p, \xi)\right) \\
&= \nabla_p\cdot\left(\int G(\xi - \bar{\xi})p - \nabla V(\xi - \bar{\xi})\, df_t^N(\bar{p}, \bar{\xi}))f_t^N(p, \xi)\right), \\
&\qquad f_0^N(p, \xi) = \hat{f}(p, \xi).
\end{aligned}$$

Note that this computation slightly differs from the standard derivation of the mean-field equation (3) as the drift with respect to ξ is shifted by the average momentum.

Remark 3. The characteristic equations of this PDE are given by

$$\begin{aligned}
\frac{d}{dt}\xi_t &= P_t - \int P_t(\bar{v})d\hat{f}(\bar{x}, \bar{v}), \qquad \xi_0(x) = x, \\
\frac{d}{dt}P_t &= (G * \varrho_t)(\xi_t)P_t - (\nabla V * \varrho_t)(\xi_t), \qquad P_0(v) = v.
\end{aligned}$$

Note that $\hat{f}$ is the initial distribution of the particles, the evolution of f over time is in terms of the characteristics as $f_t = (p_t, \xi_t)_\# \hat{f}$, i.e., the distribution of the particles at time $t > 0$ is the push forward of the initial distribution along the characteristics.

To recover the PHS structure, we need to perform a retransformation similar to the finite dimensional case. Due to the relative positions the transformation is based on the initial conditions of two independently chosen particles $z = (x, v)$ and $\bar{z} = (\bar{x}, \bar{v})$, respectively.

Let $\mathrm{law}(z) = \mathrm{law}(\bar{z}) = \hat{f}$. As before it holds $v = p$ and $\bar{v} = \bar{p}$. The initial conditions of the characteristics are given by

$$\xi_0(x) = x,\ \xi_0(\bar{x}) = \bar{x}, \qquad P_0(p) = p,\ P_0(\bar{p}) = \bar{p}.$$

For the retransformation we set

$$Q_t(q) := \xi_t(x) - \xi_t(\bar{x}), \quad Q_0(q) = x - \bar{x} = q.$$

Then we obtain

$$\frac{d}{dt}Q_t(q) = P_t(p) - P_t(\bar{p}), \quad \frac{d}{dt}Q_t(-q) = P_t(\bar{p}) - P_t(p),$$

which implies $Q_t(-q) = -Q_t(q)$. Moreover, the characteristics of the impulses read

$$\begin{aligned}
\frac{d}{dt}P_t(p) &= G(Q_t(q))P_t(p) - \nabla V(Q_t(q)), \\
\frac{d}{dt}P_t(\bar{p}) &= G(Q_t(-q))P_t(\bar{p}) - \nabla V(Q_t(-q)).
\end{aligned}$$

Our assumptions on G and U yield

$$G(Q_t(q) = G(Q_t(-q),\ \nabla V(Q_t(q)) = -\nabla V(Q_t(-q))$$

as in the finite dimensional case.

The Hamiltonian for $Z_t(z,\bar{z}) = (Q_t(q), Q_t(-q), P_t(p), P_t(\bar{p}))$ is given by

$$\mathcal{H}(Z_t(z,\bar{z})) = \frac{1}{2}|P_t(p)|^2 + |P_t(\bar{p})| + \frac{1}{2}V(Q_t(q)) + V(Q_t(\bar{q})),$$

to obtain the Banach space valued ODE in PHS form

$$\frac{d}{dt}Z_t = (J-R)\frac{\partial \mathcal{H}}{\partial Z_t}$$

supplemented with initial condition

$$Z_0 = (q_0, -q_0, p_0, \bar{p}_0) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$$

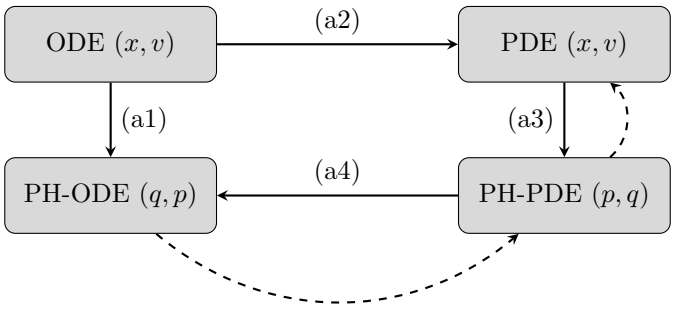
with $\text{law}(q_0) = f_q$ and $\text{law}(p_0) = \text{law}(\bar{p}_0) = f_p$. The skew-symmetric matrix J is the same as in the case $N = 2$ discussed above and

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -G(Q_t(q)) & 0 \\ 0 & 0 & 0 & -G(Q_t(-q)) \end{pmatrix}.$$

Note that the characteristic equations of the PHS systems resemble as expected. Indeed, drawing $N/2$ independent relative positions q_i and N velocities p_i and exploiting the relationships of the finite dimensional PHS allows to recover the PHS particle dynamics.

4. RELATIONSHIP OF THE DIFFERENT APPROACHES

Starting from the ODE description of interacting particle systems that is based on Newton's Second Law, we find a reformulation as finite dimensional port-Hamiltonian System (a1). This reformulation is not directly feasible for the passage to the limit $N \to \infty$, but we obtain a mean-field formulation via a shifting with respect to the center of mass of the system. This is illustrated by the dashed arrow below (a4). The classical formulation based on positions and velocities admits a the well-known mean-field equation also formulated in position and velocity variables (a2). Considering again relative positions allows us to obtain (a3). When drawing $N/2$ independent relative positions and N independent impulses and let them follow the characteristics of the mean-field PDE in PHS formulation, we recover the finite-dimensional PHS description of the interacting particle system (a4). Finally, note that all PHS formulations are based on relative positions, therefore it is only possible to recover the original (x,v) formulation on the ODE as well as on the PDE level with center of mass shifted to $0 \in \mathbb{R}^d$.



5. OUTLOOK

In future work, we plan to employ control-strategies based on the new port-Hamiltonian formulation of interacting particle systems. Moreover, the port-Hamiltonian structure will be useful in a stability analysis of general interacting particle systems.

REFERENCES

Burger, M., Pinnau, R., Totzeck, C., Tse, O., and Roth, A. (2020). Instantaneous control of interacting particle systems in the mean-field limit. *Journal of Computational Physics*, (405), 109181.

Carrillo, J., Fornasier, M., Toscani, G., and Vecil, F. (2010). Particle, kinetic, and hydrodynamic models of swarming. In T.G. Naldi G. Pareschi L. (ed.), *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*.

Cucker, F. and Smale, S. (2007). On the mathematics of emergence. *Japan J. Math.*, 2, 197–227.

D'Orsogna, M., Chuang, Y.L., Bertozzi, A.L., and Chayes, L.S. (2006). Self-propelled particles with soft-core interactions: Patterns, stability, and collapse. *Physical Review Letters*, 96(10), 104302.

Duindam, V., Macchelli, A., Stramigioli, S., and Bruyninckx, H. (eds.) (2009). *Modeling and Control of Complex Physical Systems*. Springer, Germany.

Eberard, D., Maschke, B.M., and van der Schaft, A.J. (2007). An extension of Hamiltonian systems to the thermodynamic phase space: towards a geometry of nonreversible processes. *Rep. Math. Phys.*, 60(2), 175–198. doi:10.1016/S0034-4877(07)00024-9.

Jacob, B. and Zwart, H. (2012). *Linear Port-Hamiltonian Systems on Infinite-dimensional Spaces*. Number 223 in Operator Theory: Advances and Applications. Springer, Germany.

Matei, I., Mavridis, C., Baras, J., and Zhenirovskyy, M. (2019). Inferring particle interaction physical models and their dynamical properties. *IEEE 58th Conference on Decision and Control*.

van der Schaft, A. (2006). Port-Hamiltonian systems: an introductory survey. *Proceedings on the International Congress of Mathematicians, Vol. 3, pags. 1339-1366*.

van der Schaft, A. and Maschke, B. (2002). Hamiltonian formulation of distributed-parameter systems with boundary energy flow. *Journal of Geometry and Physics*, 42, 166–194. doi:10.1016/S0393-0440(01)00083-3.

Villegas, J. (2007). *A Port-Hamiltonian Approach to Distributed Parameter Systems*. Ph.D. thesis, University of Twente, Netherlands.

Zwart, H., Le Gorrec, Y., Maschke, B., and Villegas, J. (2010). Well-posedness and regularity of hyperbolic boundary control systems on a one-dimensional spatial domain. *ESAIM: COCV*, 16(4), 1077–1093.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Optimal intervention in transportation networks

EXTENDED ABSTRACT

L. Cianfanelli * **G. Como** * **A. Ozdaglar** ** **F. Parise** ***

* *Department of Mathematical Sciences, Politecnico di Torino (e-mail: {leonardo.cianfanelli,giacomo.como}@polito.it).*
** *Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (e-mail: asuman@mit.edu)*
*** *Department of Electrical and Computer Engineering, Cornell University (e-mail: fp264@cornell.edu)*

Abstract:
We study a network design problem (NDP) where the planner aims at selecting the optimal single-link intervention in a transportation network to minimize the total congestion. Our first result is to show that the NDP may be formulated in terms of electrical quantities on a related resistor network, in particular in terms of the effective resistance between adjacent nodes. We then suggest an approach to approximate such an effective resistance by performing only local computations, and exploit this approach to design an efficient algorithm to solve the NDP, without recomputing the equilibrium flow after the intervention. We then study the optimality of the proposed procedure for recurrent networks, and provide simulations over relevant networks. *

* This extended abstract is an update of a previous work that was submitted and accepted to MTNS 2020.

Keywords: Transportation systems; Mathematical theory of networks and circuits.

1. INTRODUCTION

Congestion of transportation networks leads to massive waste of time and money (European Union (2021)). To mitigate such a problem, one approach is to optimize the underlying network (e.g., add lanes to existing roads or construct new roads) given a certain budget. This class of problems, known as *network design problem* (NDP), has been first defined in LeBlanc (1975). For a complete survey we refer to Farahani et al. (2013). NDPs are typically computationally hard. In this work we focus on a specific NDP where one link only can be improved, and show how to approximate its solution in a tractable way when the link delay functions are affine functions of the flow. Our contribution is twofold. First, we derive an alternative formulation of the problem in terms of electrical quantities, in particular the effective resistance between adjacent nodes. Then, we propose a method to locally approximate such a quantity, and based on this method we propose an efficient algorithm to find an approximated solution of the original problem. Our work is related to Steinberg and Zangwill (1983) and Dafermos and Nagurney (1984), where the travel time variation after the addition of a new route is studied with similar assumptions as ours. The main objective of those works is to investigate the sign of the travel time variation, i.e., the emergence of Braess' paradox. We instead study the problem from an optimization perspective, and quantify the travel time variation corresponding to a single link intervention. Another related problem is the optimal toll design (see, e.g., Hoefer et al. (2008) and Jelinek et al. (2014)). However, the NDP and toll design differ in the fact that tolling schemes modify the equilibrium flows on the network, but the performance of the tolls is evaluated with respect to the original delay functions. Instead, in NDP the intervention modifies both the equilibrium flows and the delay functions according to which the performance of the NDP is evaluated.

2. MODEL AND PROBLEM STATEMENT

2.1 Notation and setting

We model the transportation network as a directed multi-graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with a single origin o and destination d, where $\mathcal{N}$ and $\mathcal{E}$ denote respectively the node and the link sets. We let $\xi(e)$ and $\theta(e)$ denote respectively the head and the tail of the link e. Let $\mathcal{R}$ denote the set of the routes from o to d. Let τ denote the throughput of non-atomic agents moving from the origin o to the destination d, and $\nu = \tau(\delta_{\mathrm{o}} - \delta_{\mathrm{d}}) \in \mathbb{R}^{\mathcal{N}}$ denote the net inflow to the network, where δ_i indicates the vector with 1 in position i and 0 in the other positions. An admissible route flow is a vector $z \in \mathbb{R}^{\mathcal{R}}_+$ satisfying the constraint $z^T\mathbf{1} = \tau$. A route flow z induces a unique link flow $f \in \mathbb{R}^{\mathcal{E}}_+$ via the relation

$$f_e = \sum_{r \in \mathcal{R}: e \in r} z_r, \tag{1}$$

i.e., the flow on link e is the sum of the flow over the routes including link e. We endow each link with a delay

function, which is assumed affine, non-negative and strictly increasing, i.e., in the form

$$d_e(f_e) = a_e f_e + b_e, \quad a_e > 0, \quad b_e \geq 0, \quad \forall e \in \mathcal{E},$$

and define the cost of the route r under flow f as

$$c_r(f) = \sum_{e \in \mathcal{E}: e \in r} d_e(f_e), \tag{2}$$

which is the sum of the delay functions of the links belonging to the route.

Definition 1. An *affine routing game* is a quadruple $(\mathcal{G}, a, b, \nu)$.

A Wardrop equilibrium is a flow distribution such that no one has incentive in changing route.

Definition 2. (Wardrop equilibrium). A feasible route flow z^*, with associated link flow f^* obtained via (1), is a Wardrop equilibrium if for every route r

$$z_r^* > 0 \implies c_r(f^*) \leq c_q(f^*), \quad \forall q \in \mathcal{R}.$$

Let $B \in \mathbb{R}^{\mathcal{N} \times \mathcal{E}}$ denote the node-link incidence matrix, with entries $B_{ne} = 1$ if $n = \xi(e)$, $B_{ne} = -1$ if $n = \theta(e)$, or $B_{ne} = 0$ otherwise. It is known that a link flow f^* is a Wardrop equilibrium of a routing game if and only if it is solution of the following convex program:

$$f^* \in \underset{f \in \mathbb{R}_+^{\mathcal{E}}: Bf = \nu}{\arg\min} \quad \sum_{e \in \mathcal{E}} \int_0^{f_e} d_e(s) ds, \tag{3}$$

where $Bf = \nu$ is the projection of $z^T \mathbf{1} = \tau$ in the space of link flows. Since the delay functions are assumed strictly increasing, the objective function is strictly convex and the Wardrop equilibrium f^* is unique (Beckmann et al. (1956)). We now define the social cost, which is the total travel time at the equilibrium.

Definition 3. (Social cost). Let f^* be the unique Wardrop equilibrium of an affine routing game. The social cost is

$$C(f^*) = \sum_{e \in \mathcal{E}} f_e^* d_e(f_e^*).$$

The social cost can be interpreted as a measure of performance of the transportation network by a planner that aims at minimizing the overall congestion of the network.

2.2 Problem statement

We consider a NDP where the planner can rescale the slopes of the delay functions, from a_e to $\tilde{a}_e = a_e/(\kappa_e + 1)$, with $\kappa_e \geq 0$. We let $h_e : [0, +\infty) \to [0, +\infty)$ denote the cost associated to the intervention on link e. For every link e, we assume that h_e is non-decreasing and convex in κ_e, with $h_e(0) = 0$. The goal of the planner is to minimize a combination of the social cost and the intervention cost, where $\alpha \geq 0$ is the trade-off parameter. Specifically, by letting $f^*(\kappa)$ denote the Wardrop equilibrium corresponding to intervention κ, the NDP reads

$$\kappa^* \in \underset{\kappa \geq \mathbf{0}}{\arg\min} \sum_{e \in \mathcal{E}} \frac{a_e}{1 + \kappa_e} \left(f_e^*(\kappa)\right)^2 + b_e f_e^*(\kappa) + \alpha \sum_{e \in \mathcal{E}} h_e(\kappa_e),$$

with

$$f^*(\kappa) = \underset{f \in \mathbb{R}_+^{\mathcal{E}}: Bf = \nu}{\arg\min} \quad \sum_{e \in \mathcal{E}} \frac{a_e}{2(1 + \kappa_e)} f_e^2 + b_e f_e \tag{4}$$

This problem is in general non-convex, and hard to solve because of its bi-level nature, in the sense that the planner optimizes the network intervention κ, but the cost function depends on κ also via the Wardrop equilibrium $f^*(\kappa)$, which in turn is solution of the optimization problem (4), whose objective function depends on the intervention κ itself. For these reasons, we restrict our analysis to the case where the planner can intervene on a single link. For this special class of interventions, we are able to rephrase the problem into a single-level optimization problem and to provide an electrical interpretation of the problem. We express interventions as pairs $(e, \kappa) \in \mathcal{I}$, where κ is now a scalar value and $\mathcal{I}$ denotes the set of the feasible interventions. Let $(\mathcal{G}, a(e, \kappa), b, \nu)$ and $f^*(e, \kappa)$ denote the modified routing game and the corresponding Wardrop equilibrium respectively, and $\Delta C(e, \kappa) = C(f^*) - C_{e,\kappa}(f^*(e, \kappa))$ denote the corresponding social cost gain. Our goal is to identify the optimal intervention (e^*, κ^*). The problem can be expressed as follows.

Problem 1. Let $(\mathcal{G}, a, b, \nu)$ be an affine routing game and $\alpha \geq 0$ a trade-off parameter. Find

$$(e^*, \kappa^*) \in \underset{(e,\kappa) \in \mathcal{I}}{\arg\max} \quad \Delta C(e, \kappa) - \alpha h_e(\kappa).$$

3. RESULTS

3.1 Electrical formulation

As anticipated, the social cost variation in the transportation network may be formulated in terms of electrical quantities on a related resistor network. Let us define how to construct such a resistor network $\mathcal{G}_R$.

Definition 4. Given the transportation network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, we construct the *associated resistor network* $\mathcal{G}_R = (\mathcal{N}, \mathcal{L}, W)$ as follows:

- the node set $\mathcal{N}$ is equivalent.
- $W \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ is the conductance matrix, which reads

$$W_{ij} := \begin{cases} \displaystyle\sum_{\substack{e \in \mathcal{E}: \\ \xi(e)=i, \theta(e)=j, \text{ or} \\ \xi(e)=j, \theta(e)=i}} \frac{1}{a_e} & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases} \tag{5}$$

  Note that W is symmetric, thus $\mathcal{G}_R$ is undirected. The element W_{ij} indicates the conductance between nodes i and j.
- Multiple links from the same pair of nodes are not allowed, so that every link $l \in \mathcal{L}$ can be identified by an unordered pair of nodes $\{i, j\}$, and the set $\mathcal{L}$ is univocally determined by non-zero elements of W. We define a mapping $M : \mathcal{E} \to \mathcal{L}$ that associates to every link $e \in \mathcal{E}$ of the transportation network the corresponding link $l = M(e) = \{\xi(e), \theta(e)\}$ of the resistor network.

Definition 5. (Effective resistance). Consider a resistor network $(\mathcal{N}, \mathcal{L}, W)$ and an arbitrary link $l = \{i, j\} \in \mathcal{L}$. We define the effective resistance r_l of the link l as the effective resistance between i and j, i.e.,

$$r_l = \overline{v}_i - \overline{v}_j,$$

where $\overline{v}$ is the voltage vector when unitary current from i to j is injected, i.e.,

$$\sum_k W_{hk}(\overline{v}_h - \overline{v}_k) = \delta_i - \delta_j \qquad \forall h \in \mathcal{N}. \tag{6}$$

The next theorem establishes a relation between the social cost variation corresponding to a single-link interventions on the transportation network and the associated resistor network. The result holds under the following regularity assumption, which states that all the feasible interventions do not modify the support of the Wardrop equilibrium.

Assumption 1. Let $\mathcal{E}_+(e,\kappa)$ be the set of links $j \in \mathcal{E}$ such that $f_j^*(e,\kappa) > 0$ in the Wardrop equilibrium of $(\mathcal{G}, a(e,\kappa), b, \nu)$. We assume that $\mathcal{E}_+(e,\kappa) = \mathcal{E}$ for every $(e,\kappa) \in \mathcal{I}$.

Theorem 1. Let $(\mathcal{G}, a, b, \nu)$ be a routing game, and suppose Assumption 1 holds. Then,

$$\Delta C(e,\kappa) = \iota \frac{f_e^*(v_{\xi(e)} - v_{\theta(e)})}{\frac{1}{\kappa} + \frac{r_{M(e)}}{a_e}}, \tag{7}$$

where v is the voltage vector over the associated resistor network when a unitary current from o to d is injected, $r_{M(e)}$ is the effective resistance of the link $M(e)$, and ι is a positive constant.

The idea behind the proof is that with affine delay function the Wardrop equilibrium is the result of a quadratic program with constraints, whose KKT conditions are linear. Moreover, it is possible to relate the social cost variation to Lagrangian multipliers. Since Assumption 1 guarantees that interventions are rank-1 perturbation of such linear system, we exploit Sherman-Morrison relation to compute the Lagrangian multiplier variation, and then relate it to electrical quantities by using some electrical circuits theory. Observe that the social cost variation is expressed in terms of quantities that do not depend on the intervention, and does not require the computation of the new Wardrop equilibrium. In particular, f^* can be computed by solving the convex program (3), and the voltage v by solving a sparse linear system analogous to (6). Also the constant ι can be easily computed. The computational bottleneck is given by the computation of all the link effective resistances, which require to solve $|\mathcal{L}|$ sparse linear systems. In the next section we propose a method to approximate r_l that does not scale with the size of the network. Before doing that, we discuss Assumption 1. The assumption is not new in NDP literature (Steinberg and Zangwill (1983) and Dafermos and Nagurney (1984)). Moreover, the next lemma shows that it is without loss of generality on series-parallel graphs, provided that the throughput is sufficiently large.

Lemma 1. Let $(\mathcal{G}, a, b, \nu)$ be a routing game, and assume that $\mathcal{G}$ is series-parallel. Then, there exists $\overline{\tau}$ such that for every $\tau \geq \overline{\tau}$ the set of links such that $f_e^* > 0$ does not depend on a.

The result above follows from the fact that along all routes of series-parallel networks the nodes are ordered in such a way that the Lagrangian multipliers corresponding to relation $Bf = \nu$ in (3) are always decreasing from the origin to the destination.

3.2 Our algorithm

In this section we propose an algorithm to solve in approximation Problem 1. Our algorithm relies on Theorem 1, and on the idea that the effective resistance of a link can be approximated by looking at a local portion of the network. Let us introduce the following operations.

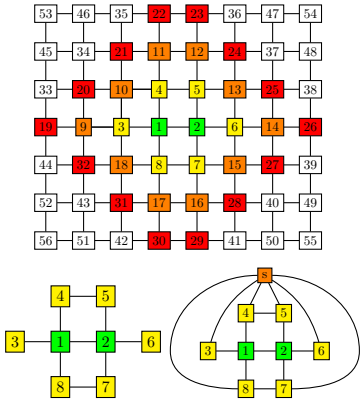


Fig. 1. Square grid. *Above*: the yellow, orange and red nodes are at distance 1, 2 and 3, respectively from the green nodes. *Bottom left*: cut at distance 1. *Bottom right*: shorted at distance 1.

Definition 6. A resistor network $\mathcal{G}_R$ is cut at distance d with respect to a link $l = \{i,j\} \in \mathcal{L}$ if every node at distance greater than d from link l, i.e., from both i and j, is removed, and every link having at least one endpoint in the set of the removed nodes is removed. Let $\mathcal{G}_l^{U_d}$ and $r_l^{U_d}$ denote such a network and the effective resistance of link l on it, respectively.

Definition 7. A resistor network $\mathcal{G}_R$ is shorted at distance d with respect to a link $l \in \mathcal{L}$ if all the nodes at distance greater than d from the link l are shorted together, i.e., an infinite conductance is added between each pair of such nodes. Let $\mathcal{G}_l^{L_d}$ and $r_l^{L_d}$ denote such a network and the effective resistance of link l on it, respectively.

We refer to Fig. 1 for an example of cut and shorted regular grid. The next proposition states that cutting and shorting a network provide respectively upper and lower bounds for the effective resistance of a link. Moreover, the tightness of the bounds is a monotone function of the distance d.

Proposition 1. Let $\mathcal{G}_R$ be a resistor network. Then, for every link $l = \{i,j\} \in \mathcal{L}$,

$$r_l^{U_{d_1}} \geq r_l^{U_{d_2}} \geq r_l \geq r_l^{L_{d_2}} \geq r_l^{L_{d_1}}, \quad \forall\ d_2 > d_1 \geq 1.$$

The following algorithm provides an approximated solution of Problem 1 based on Proposition 1.

Theorem 2. Let $\Delta C(e,\kappa)$ be the cost variation corresponding to intervention (e,κ) as given in Theorem 1, and

$$\Delta C_d(e,\kappa) = \iota \frac{f_e^*(v_{\xi(e)} - v_{\theta(e)})}{\frac{1}{\kappa} + \frac{r_{M(e)}^{U_d} + r_{M(e)}^{L_d}}{2a_e}}$$

be the cost variation estimated by Algorithm 1 for a given distance $d \geq 1$. Then,

$$\left| \frac{\Delta C(e,\kappa) - \Delta C_d(e,\kappa)}{\Delta C(e,\kappa)} \right| \leq \frac{\epsilon_{ed}}{2\left(\frac{1}{\kappa} + \frac{r_{M(e)}^{U_d} + r_{M(e)}^{L_d}}{2a_e}\right)}$$

where

$$\epsilon_{ed} := \frac{r_{M(e)}^{U_d} - r_{M(e)}^{L_d}}{a_e}.$$

Furthermore,

$$\Delta C(e,\kappa) \geq \iota \frac{f_e^*(v_{\xi(e)} - v_{\theta(e)})}{\frac{1}{\kappa} + \frac{r_{M(e)}^{U_d}}{a_e}}. \tag{8}$$

Algorithm 1:

Input: The affine routing game $(\mathcal{G}, a, b, \nu)$, the cost functions $\{h_e\}_{e\in\mathcal{E}}$, and the distance $d \geq 1$ used to approximate the effective resistance.
Output: The optimal intervention $(e,\kappa)^{*d} \in \mathcal{I}$.
Construct the associated resistor network $\mathcal{G}_R$.
Compute v by solving the sparse linear system

$$\sum_k W_{hk}(v_h - v_k) = \delta_o - \delta_d \qquad \forall h \in \mathcal{N}.$$

for *each* $l \in \mathcal{L}$ **do**
  Construct $\mathcal{G}_l^{U_d}$ and $\mathcal{G}_l^{L_d}$;
  Compute $r_l^{U_d}$ on $\mathcal{G}_l^{U_d}$, and $r_l^{L_d}$ on $\mathcal{G}_l^{L_d}$.
end
for *each* $e \in \mathcal{E}$ **do**
  Select κ_e^{*d} such that

$$\kappa_e^{*d} \in \arg\max_{\kappa:(e,\kappa)\in\mathcal{I}} \quad \iota \frac{f_e^*(v_{\xi(e)} - v_{\theta(e)})}{\frac{1}{\kappa} + \frac{r_{M(e)}^{U_d} + r_{M(e)}^{L_d}}{2a_e}} - \alpha h_e(\kappa).$$

end
Select $(e,\kappa)^{*d}$ such that

$$(e,\kappa)^{*d} \in \arg\max_{(e,\kappa_e^{*d})} \quad \iota \frac{f_e^*(v_{\xi(e)} - v_{\theta(e)})}{\frac{1}{\kappa} + \frac{r_{M(e)}^{U_d} + r_{M(e)}^{L_d}}{2a_e}} - \alpha h_e(\kappa).$$

Remark 1. Observe that the upper and lower bounds of the effective resistance of a link depend only on the local structure of the network. If we assume that the local structure of the transportation network does not depend on its size (which is a reasonable assumption, think for instance of bidimensional grids), both the tightness of the bounds and their computational complexity do not scale with the size of the network. Thus, the complexity for computing all the links effective resistance scale linearly with the number of links (or nodes) of the network.

In the final part of the section we provide a sufficient condition on the network topology under which the gap between the upper and the lower bound of the effective resistances, and therefore ϵ_{ed} defined in Theorem 2, vanishes in the limit of infinite distance. To this end, we interpret the conductance matrix W of a resistor network as the transition rates matrix of a continuous-time random walk, and define the class of recurrent networks.

Definition 8. A network $\mathcal{G}_R = (\mathcal{N}, \mathcal{L}, W)$ is *recurrent* if the random walk with transition rates W visits its starting node infinitely often with probability one.

The next theorem states that the gap between the upper and the lower bound vanishes asymptotically on recurrent networks, provided that the degree of every node is bounded.

Theorem 3. Let $\mathcal{G}_R = (\mathcal{N}, \mathcal{L}, W)$ be an infinite recurrent resistor network, and let the weighted degree of every node be finite. Then, for every link $l \in \mathcal{L}$,

$$\lim_{d\to+\infty} (r_l^{U_d} - r_l^{L_d}) = 0.$$

Theorem 3 is a consequence of the relation between effective resistance and random walks over networks. The result states that the gap between the bounds vanish asymptotically if the associated resistor network is recurrent. In the next subsection we provide numerical results on an infinite square grid, showing that gap between the bounds is quite small even for small distances d.

3.3 Simulations on infinite grids

Grids are a significant test for our algorithm, since they are a good proxy for the transportation network of many cities. In Table 1 the performances of the upper and lower bounds for a infinite square grid are shown. Despite the grid being infinite, we get a good approximation even for small distance. The simulations show that for every link l

$$r_l^{U_d} - r_l = r_l - r_l^{L_d} = O(1/d^2).$$

The tightness of the bounds scales similarly in all the regular grids.

Table 1. Table of upper and lower bound in infinite square grid.

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
$(r_l^{U_d} - r_l)/r_l$	1/5	0.0804	0.0426	0.0262	0.0178
$(r_l - r_l^{L_d})/r_l$	1/5	0.0804	0.0426	0.0262	0.0178

4. CONCLUSION

In this work we reformulate a network design problem in terms of electrical quantities. We then propose a method to approximate the effective resistance between two adjacent nodes, and exploit this method to construct an efficient algorithm that finds in approximation the optimal intervention on the transportation network. Future research lines include the relaxation of some restrictive assumptions, as well as studying the case of multiple interventions.

REFERENCES

Beckmann, M., McGuire, C.B., and Winsten, C.B. (1956). Studies in the economics of transportation. Technical report.

Dafermos, S. and Nagurney, A. (1984). On some traffic equilibrium theory paradoxes. *Transportation Research Part B: Methodological*, 18(2), 101–110.

European Union (2021). Urban mobility. `https://transport.ec.europa.eu`. [Online; accessed 10-Jul-2022].

Farahani, R.Z., Miandoabchi, E., Szeto, W.Y., and Rashidi, H. (2013). A review of urban transportation network design problems. *European Journal of Operational Research*, 229(2), 281–302.

Hoefer, M., Olbrich, L., and Skopalik, A. (2008). Taxing subnetworks. In *International Workshop on Internet and Network Economics*, 286–294. Springer.

Jelinek, T., Klaas, M., and Schäfer, G. (2014). Computing optimal tolls with arc restrictions and heterogeneous players. In *31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

LeBlanc, L.J. (1975). An algorithm for the discrete network design problem. *Transportation Science*, 9(3), 183–199.

Steinberg, R. and Zangwill, W.I. (1983). The prevalence of Braess' paradox. *Transportation Science*, 17(3), 301–318.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Content Rendering for Acoustic Levitation Displays via Optimal Path Following

Viktorija Paneva [*] **Arthur Fleig** [*] **Diego Martínez Plasencia** [**]
Timm Faulwasser [***] **Jörg Müller** [*]

[*] *Department of Computer Science, University of Bayreuth, DE 95447 (e-mail: vpaneva@acm.org, arthur.fleig@uni-bayreuth.de, joerg.mueller@uni-bayreuth.de).*
[**] *Computer Science, University College London, UK WC1E 6BT (e-mail: d.plasencia@ucl.ac.uk)*
[***] *Department of Electrical Engineering and Information Technology, TU Dortmund University, DE 44227 (e-mail: timm.faulwasser@ieee.org).*

Abstract: Recently, volumetric displays based on acoustic levitation have demonstrated the capability to produce mid-air content using the Persistence of Vision (PoV) effect. In these displays, acoustic traps are used to rapidly move a small levitated particle along a prescribed path. This note is based on our recent work *OptiTrap* (Paneva et al., 2022), the first structured numerical approach for computing trap positions and timings via optimal control to produce feasible and (nearly) time-optimal trajectories that reveal generic levitated graphics. While previously, feasible trap trajectories needed to be tuned manually for each shape and levitator, relying on trial and error, *OptiTrap* automates this process by allowing for a systematic exploration of the range of contents that a given levitation display can render. This represents a crucial milestone for future content authoring tools for acoustic levitation displays and advances volumetric displays closer toward real-world applications.

Keywords: Ultrasonic levitation, Minimum time problems, Path following, Optimal control

1. INTRODUCTION

Acoustic levitation displays use ultrasonic waves to trap small particles in mid-air, acting as volumetric pixels (voxels). Several practical aspects have been investigated around these displays, such as low-latency particle manipulation (Bachynskyi et al., 2018), and content detection and initialisation (Fender et al., 2021).

The ability to move single (Hirayama et al., 2019) or multiple (Plasencia et al., 2020) levitated particles at very high speeds was instrumental for achieving dynamic and free-form volumetric content. However, this was nonetheless limited to relatively small sizes and simple vector graphics (Fushimi et al., 2020). Little effort was made towards optimising the levitated visual content while considering the system dynamics of such displays, particularly for challenging content such as the one created by levitated particles moving at PoV speeds. Paneva et al. (2020) proposed an interactive simulation of a levitation interface, using a model of the particle movement in such a display. The application operates in a feed-forward manner, simulating the dynamics of the particle given a specific path for the traps, however, it does not address the inverse problem.

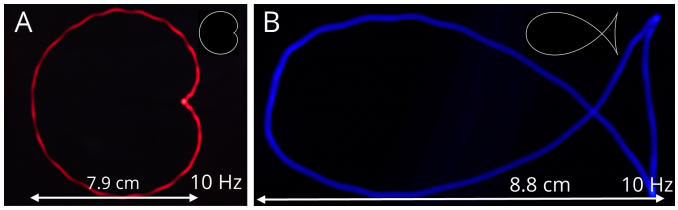


Fig. 1. Shapes involving sharp edges and significant changes in curvature demonstrated using acoustic levitation.

OptiTrap (Paneva et al., 2022) is the first algorithm allowing the definition of generic PoV content, requiring only a geometric definition (i.e., shape to present, no timing information) and optimising it according to the capabilities of the device and the dynamics of the trap-particle system. *OptiTrap* automates the definition of levitated PoV content, computing physically feasible and nearly time-optimal trap trajectories given only a reference path. This allows for larger shapes than previously demonstrated, as well as shapes featuring significant changes in curvature and/or sharp corners (Figure 1).

2. OPTIMAL CONTROL FOR LEVITATION DISPLAYS

2.1 Hardware Setup

Our setup, illustrated in Figure 2, consists of two arrays of 16×16 transducers facing each other, controlled by an FPGA, and an OptiTrack [1] tracking system - Prime 13 motion capture cameras operating at a frequency of 240Hz.

[1] www.optitrack.com

With an update rate of up to 10kHz, the device can create a single twin-trap within the levitation volume using the method from Hirayama et al. (2019). We experimentally determined the vertical and horizontal forces exerted on the particle and found that the vertical force is double the horizontal. This difference needs to be reflected by the model that is introduced in the next section.

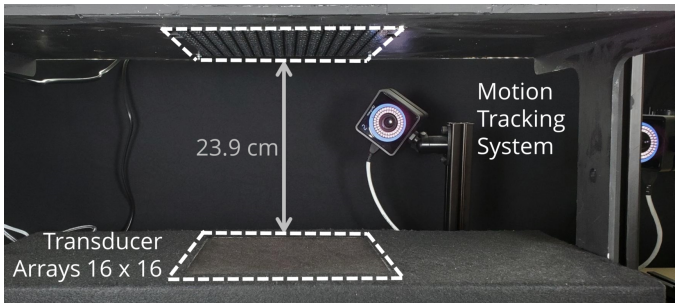


Fig. 2. Overview of the components of our levitation display.

2.2 Modelling the Trap-Particle Dynamics

To model the trap-particle dynamics, we use simple Newtonian mechanics, i.e.,

$$m\ddot{\mathbf{p}}(t) = F(\mathbf{p}(t), \dot{\mathbf{p}}(t), \mathbf{u}(t)), \tag{1}$$

where $\mathbf{p}(t) = (p_x(t), p_y(t), p_z(t))^\top \in \mathbb{R}^3$ represents the particle position in Cartesian (x, y, z) coordinates at time $t \in \mathbb{R}_0^+$, m is the particle mass, $\mathbf{u}(t) = (u_x(t), u_y(t), u_z(t))^\top \in \mathbb{R}^3$ is the control input specifying the position of the acoustic trap, and F is the net force acting on the particle. Following Hirayama et al. (2019), we neglect drag and gravitational forces due to the dominating acoustic radiation forces.

In many cases, the acoustic force can be described by the gradient of the Gor'kov potential (Bruus, 2012). Our specific setup (top-bottom transducer placement and vertical twin traps) and objective (find where to place the acoustic trap to produce a specific force) allows to consider only a region around the peak forces. In this region, the forces in our setup distribute mostly axis-symmetrically. Hence, we approximate the force as

$$F(\mathbf{p}, \mathbf{u}) = \begin{pmatrix} F_x(\mathbf{p}, \mathbf{u}) \\ F_y(\mathbf{p}, \mathbf{u}) \\ F_z(\mathbf{p}, \mathbf{u}) \end{pmatrix} := \begin{pmatrix} F_r(\mathbf{p}, \mathbf{u}) \cos\phi \\ F_r(\mathbf{p}, \mathbf{u}) \sin\phi \\ F_z(\mathbf{p}, \mathbf{u}) \end{pmatrix}, \tag{2}$$

where

$$F_r(\mathbf{p}, \mathbf{u}) := \mathcal{A}_r \cdot \cos\left(\mathcal{V}_z \cdot (u_z - p_z)\right) \cdot \sin\left(\mathcal{V}_{xr} \cdot \sqrt{(u_x - p_x)^2 + (u_y - p_y)^2}\right), \tag{3a}$$

$$F_z(\mathbf{p}, \mathbf{u}) := \mathcal{A}_z \cdot \sin\left(\mathcal{V}_z \cdot (u_z - p_z)\right) \cdot \cos\left(\mathcal{V}_{zr} \cdot \sqrt{(u_x - p_x)^2 + (u_y - p_y)^2}\right), \tag{3b}$$

$$\phi = \arctan\left(\frac{u_y - p_y}{u_x - p_x}\right),$$

and where $\mathcal{A}_r, \mathcal{A}_z$ denote the peak forces along the radial and vertical directions of the trap, respectively, and $\mathcal{V}_z$, $\mathcal{V}_{xr}$, $\mathcal{V}_{zr}$ are the characteristic frequencies of the sinusoidals describing how the forces evolve around the trap. For more details on this approximation procedure we refer to the original paper (Paneva et al., 2022).

2.3 Rendering Content via Path Following

The task at hand is to render, as fast as possible, arbitrary complex objects, formulated as an explicitly parameterised curve

$$Q := \{\xi \in \mathbb{R}^3 \mid \theta \in [\theta_0, \theta_f] \mapsto \mathbf{q}(\theta)\}, \tag{4}$$

where we require $\mathbf{q} \in C^2(\mathbb{R}; \mathbb{R}^3)$. The path parameter θ models the progress on the path from the starting point $\mathbf{q}(\theta_0)$ to the end point $\mathbf{q}(\theta_f)$. To create the PoV effect, we consider periodic paths, for which $\mathbf{q}(\theta_0) = \mathbf{q}(\theta_f)$ and $\dot{\mathbf{q}}(\theta_0) = \dot{\mathbf{q}}(\theta_f)$ hold. For example, consider the shape of the cardioid in Figure 1 (left), which can be described by $\mathbf{q}(\theta) = (0, r\sin(\theta)(1+\cos(\theta)), -r\cos(\theta)(1+\cos(\theta)) + r)^\top$, where $\theta \in [0, 2\pi]$ and $r > 0$.

Since Q comes without any preassigned time information, we need to determine the timing $t \mapsto \theta(t)$. Following Faulwasser (2012); Faulwasser et al. (2017), we assume the particle follows the path Q exactly at all times, i.e., $\mathbf{p}(t) - \mathbf{q}(\theta(t)) \equiv 0$. This leads to

$$\mathbf{p}(t) = \mathbf{q}(\theta(t)), \tag{5a}$$

$$\dot{\mathbf{p}}(t) = \dot{\mathbf{q}}(\theta(t)) = \frac{\partial \mathbf{q}}{\partial \theta}\dot{\theta}(t), \tag{5b}$$

$$\ddot{\mathbf{p}}(t) = \ddot{\mathbf{q}}(\theta(t)) = \frac{\partial^2 \mathbf{q}}{\partial \theta^2}\dot{\theta}(t)^2 + \frac{\partial \mathbf{q}}{\partial \theta}\ddot{\theta}(t). \tag{5c}$$

Using a virtual function $v(t) \in \mathbb{R}$ to control the progress of the particle along Q, the timing law is modelled as a double integrator

$$\ddot{\theta}(t) = v(t) \tag{6}$$

to avoid large jumps in the acceleration. To keep the periodic nature, we impose

$$\theta(0) = \theta_0, \quad \theta(T) = \theta_f, \quad \dot{\theta}(0) = \dot{\theta}(T), \tag{7}$$

where the traversal time T will be an optimisation variable in the latter optimal control problem (12). Using

$$\mathbf{z}(t) := (\theta(t), \dot{\theta}(t))^\top$$

we rewrite (6)-(7) as

$$\dot{\mathbf{z}}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \mathbf{z}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} v(t), \ \mathbf{z}(0) = \mathbf{z}_0, \ \mathbf{z}(T) = \mathbf{z}_T, \tag{8}$$

solvable with standard Runge-Kutta methods (Butcher, 2016).

2.4 Coupling Path Following with Trap-Particle Dynamics

To render content on the levitator, we couple the path-following approach from Section 2.3 with the trap-particle dynamics (1)-(3) from Section 2.2. The usual approach of rewriting (1) as

$$M(\ddot{\mathbf{p}}(t), \mathbf{p}(t), \mathbf{u}(t)) := m\ddot{\mathbf{p}}(t) - F(\mathbf{p}(t), \mathbf{u}(t)) = 0 \tag{9}$$

and finding a local inversion $\mathbf{u}(t) = M^{-1}(\ddot{\mathbf{p}}(t), \mathbf{p}(t))$ is not straightforward for (3). Hence, we tackle this task via numerical optimisation by introducing auxiliary variables

$$\zeta_1 := \sin(\mathcal{V}_{xr}\sqrt{(u_x - q_x(\theta))^2 + (u_y - q_y(\theta))^2}), \quad (10a)$$
$$\zeta_2 := \cos \mathcal{V}_z \cdot (u_z - q_z(\theta)), \quad (10b)$$
$$\zeta_3 := \sin \mathcal{V}_z \cdot (u_z - q_z(\theta)), \quad (10c)$$
$$\zeta_4 := \cos \mathcal{V}_{zr}\sqrt{(u_x - q_x(\theta))^2 + (u_y - q_y(\theta))^2}), \quad (10d)$$
$$\zeta_5 := \sin \phi, \quad (10e)$$
$$\zeta_6 := \cos \phi, \quad (10f)$$

for each trigonometric term in (3), where p_i is replaced by $q_i(\theta)$, $i \in \{x, y, z\}$. This allows us to formally express (2) in terms of $\zeta := (\zeta_1, ..., \zeta_6)$:

$$\tilde{F}(\zeta) := \begin{pmatrix} \mathcal{A}_r \zeta_1 \zeta_2 \zeta_6 \\ \mathcal{A}_r \zeta_1 \zeta_2 \zeta_5 \\ \mathcal{A}_z \zeta_4 \zeta_3 \end{pmatrix}.$$

Similar to (9), along the path Q we define

$$\widetilde{M}(\mathbf{z}(t), v(t), \zeta(t)) := m\ddot{\mathbf{q}}(\theta(t)) - \tilde{F}(\zeta(t)) = 0. \quad (11)$$

With this approach, we will need to extract the trap positions $\mathbf{u}(t)$ by solving (10) numerically. To counter numerical instabilities that could occur in particular for ζ_i approaching ± 1, we introduce additional constraints

$$\mathcal{Z} := \left\{\zeta \in [\varepsilon - 1, 1 - \varepsilon]^6 \mid \zeta_2^2 + \zeta_3^2 = 1,\ \zeta_5^2 + \zeta_6^2 = 1\right\},$$

with a user-chosen back-off parameter $\varepsilon \in]0, 1[$ that captures the trade-off between numerically "stable" solutions for $\mathbf{u}(t)$ and exploiting the maximum forces of the device.

The final OCP is then given by

$$\begin{aligned} \min_{v,T,\zeta}\ & T + \gamma \int_0^T v(t)^2 \mathrm{d}t \\ \text{subject to } & (8), \\ & \widetilde{M}(\mathbf{z}(t), v(t), \zeta(t)) = 0, \\ & \zeta(t) \in \mathcal{Z}. \end{aligned} \quad (12)$$

Lastly, we discretised the final OCP and solved the resulting nonlinear programming problem using Ipopt (Andersson et al., 2018). The function evaluations and the computation of the derivatives were performed with CasADi (Wächter and Biegler, 2006).

3. EVALUATION

Next we compare our *OptiTrap* approach against a *Baseline*, where the path parameter is homogeneously sampled and the traps are placed directly on the reference path of the particle. The evaluation was conducted on four test shapes: circle, cardioid, squircle and fish, as shown in Figures 3 and 4. The particle motion shown in these figures comes not from simulations, but from actual experimental results, captured using the OptiTrack tracking system.

In the first part of the evaluation, we investigate the maximum shape size that can be rendered in PoV time (0.1s) with each approach. We see in Figure 3 that the most striking difference was obtained for the squircle, where with *OptiTrap* 1.9 meters of content per second was rendered, and with the *Baseline* only 0.29. The increase in size was similar for both the cardioid (2.80m of content per s with *OptiTrap*, 2.48 with the *Baseline*) and the fish (2.75m of content per s with *OptiTrap*, 2.42 with the *Baseline*), while there was not significant difference for the circle. This is not surprising, as the circle is the simplest and most homogeneous shape in the test parkour.

In the second part of the evaluation, we investigate the maximum possible rendering frequency with *OptiTrap* and the *Baseline*, while keeping the size of the test shape constant. The results are illustrated in Figure 4. As expected,

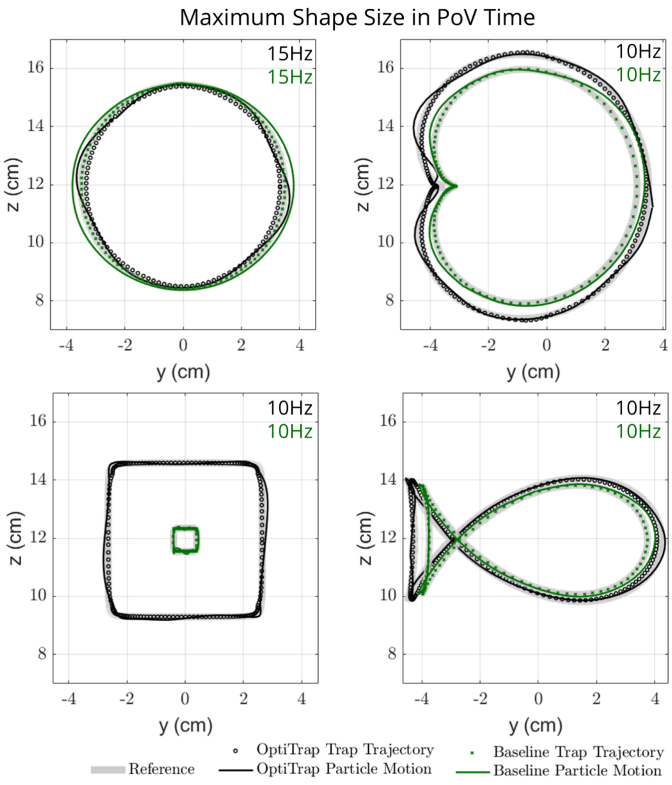


Fig. 3. Maximum shape size of *OptiTrap* vs. the Baseline.

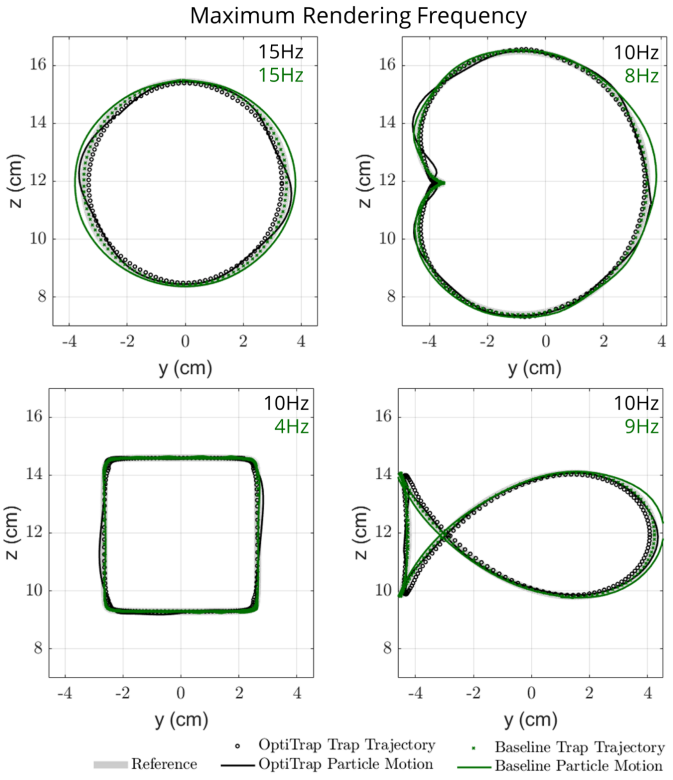


Fig. 4. Maximum rendering frequency of *OptiTrap* vs. the Baseline.

the same rendering frequency was obtained for the circle using both methods. However, the cardioid was rendered at 10Hz using *OptiTrap*, and 8Hz with the *Baseline* (25% increase). The rendering frequency for the fish increased by 11% using *OptiTrap*, i.e., from 9 to 10Hz, and lastly, we obtain an increase of 150% for the squircle.

For a more extensive and detailed evaluation of *OptiTrap*, also including a comparison to a more sophisticated baseline, please refer to Paneva et al. (2022).

4. CONCLUSION

We briefly discussed *OptiTrap* – a structured numerical approach to compute trap trajectories for acoustic levitation displays. *OptiTrap* automatically computes physically feasible and nearly time-optimal trap trajectories to reveal generic levitated content in mid-air, assuming only a reference path. This is a particularly important step for the adoption of PoV levitation displays, as it allows the content designers to focus on the shapes to be rendered, with feasible solutions taking into account the capabilities of the specific device, being computed automatically by the algorithm. As such, *OptiTrap* has the potential to become an instrumental tool in helping to further explore and develop these displays. In the future, this method can be extended to include visual content composed of multiple levitated particles, it can be applied to other domains, such as photophoretic displays or containerless matter transportation for applications in pharmacy and biochemistry, or can be used as a base for developing more complex learning-based approaches.

ACKNOWLEDGEMENTS

This project was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement #737087 (Levitate)

REFERENCES

Andersson, J.A.E., Gillis, J., Horn, G., Rawlings, J.B., and Diehl, M. (2018). CasADi – A software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*.

Bachynskyi, M., Paneva, V., and Müller, J. (2018). Levicursor: Dexterous interaction with a levitating object. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, ISS '18, 253–262. ACM, New York, NY, USA. doi:10.1145/3279778.3279802. URL `http://doi.acm.org/10.1145/3279778.3279802`.

Bruus, H. (2012). Acoustofluidics 7: The acoustic radiation force on small particles. *Lab Chip*, 12, 1014–1021. doi:10.1039/C2LC21068A.

Butcher, J.C. (2016). *Numerical methods for ordinary differential equations*.

Faulwasser, T. (2012). *Optimization-based solutions to constrained trajectory-tracking and path-following problems*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, Universitaetspl. 2.

Faulwasser, T., Weber, T., Zometa, P., and Findeisen, R. (2017). Implementation of nonlinear model predictive path-following control for an industrial robot. *IEEE Trans. Contr. Syst. Techn.*, 25(4), 1505–1511. doi:10.1109/TCST.2016.2601624.

Fender, A.R., Martinez Plasencia, D., and Subramanian, S. (2021). *ArticuLev: An Integrated Self-Assembly Pipeline for Articulated Multi-Bead Levitation Primitives*. Association for Computing Machinery, New York, NY, USA. URL `https://doi.org/10.1145/3411764.3445342`.

Fushimi, T., Drinkwater, B.W., and Hill, T.L. (2020). What is the ultimate capability of acoustophoretic volumetric displays? *Applied Physics Letters*, 116(24), 244101. doi:10.1063/5.0008351. URL `https://doi.org/10.1063/5.0008351`.

Hirayama, R., Plasencia, D.M., Masuda, N., and Subramanian, S. (2019). A volumetric display for visual, tactile and audio presentation using acoustic trapping. *Nature*, 575(7782), 320–323.

Paneva, V., Bachynskyi, M., and Müller, J. (2020). Levitation simulator: Prototyping for ultrasonic levitation interfaces in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. ACM, New York, NY, USA.

Paneva, V., Fleig, A., Plasencia, D.M., Faulwasser, T., and Müller, J. (2022). Optitrap: Optimal trap trajectories for acoustic levitation displays. *ACM Trans. Graph. (In Press)*. doi:10.1145/3517746. URL `http://dx.doi.org/10.1145/3517746`.

Plasencia, D.M., Hirayama, R., Montano-Murillo, R., and Subramanian, S. (2020). Gs-pat: High-speed multi-point sound-fields for phased arrays of transducers. *ACM Trans. Graph.*, 39(4). doi:10.1145/3386569.3392492.

Wächter, A. and Biegler, L.T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1), 25–57.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Sparsity Enforcing Convex Relaxation of D-Optimal Sensor Selection for Parameter Estimation of Distributed Parameter Systems

Dariusz Uciński

Institute of Control and Computation Engineering
University of Zielona Góra, Zielona Góra, Poland
(e-mail: d.ucinski@issi.uz.zgora.pl)

Abstract: Owing to the wide availability of efficient convex optimization algorithms, convex relaxation of optimum sensor selection problems has gained in popularity. Generally, however, there is a performance gap between the optimal solution of the original combinatorial problem and the heuristic solution of the respective relaxed continuous problem. This gap can be small in many cases, but there is no guarantee that this is always the case. That is why the D-optimality criterion is often extended by addition of some kind of sparsity-enforcing penalty term. Unfortunately, the problem convexity is then lost and the question of how to control the influence of this penalty so as not to excessively deteriorate the optimal relaxed solution remains open. This work proposes an alternative problem formulation, in which the sparsity-promoting term is directly minimized subject to the constraint that the D-efficiency of the sensor selection is no less than a given threshold. This offers direct control of the degree of optimality of the produced solution. An efficient computational scheme based on the majorization-minimization algorithm is proposed, which reduces to solving a sequence of low-dimensional convex optimization problems via generalized simplicial decomposition. A numerical example illustrating the effectiveness of the proposed approach is also reported.

Keywords: system identification, distributed parameter systems, sensor location, convex relaxation, sparsification.

MSC 2010: 62K05, 35Q93, 90C25.

1. INTRODUCTION

Distributed parameter systems (DPSs) constitute a class of dynamic systems whose states depend not only on time, but also on the spatial variable. Their adequate description are partial differential equations (PDEs). In most cases, not all physical parameters underlying such models can be directly measured and they have to be estimated via calibration yielding the best fit of the model output to the observations of the actual system which are provided by measurement sensors. But the number of sensors is usually limited and practitioners face the problem of where to locate them so as to collect the most valuable information about the parameters.

The traditional approach to optimal sensor location consists in formulating it in terms of an optimization problem employing various design criteria defined on the Fisher information matrix (FIM) associated with the estimated parameters. Comprehensive overviews of the works published in this area are contained in the monographs by Uciński (2005), Patan (2012) and Rafajłowicz (2022).

In the past two decades, the interest in this problem has increased rapidly due to the growing popularity of sensor networks. More and more complex scenarios have been investigated to accomodate to more and more demanding practical settings. They have primarily been concentrated around properly addressing the ill-posedness inherent in problems with large (or even infinite) dimensions of the parameter space, see works by Alexanderian (2021), Alexanderian et al. (2014), Gejadze and Shutyaev (2012) and Haber et al. (2010).

The number of sensors is usually fixed and imposed by the available experimental budget. Most techniques come down to the selection of optimal sensor locations from a finite (but possibly very large) set of candidate locations. Note that the problem of assigning sensors to specific spatial locations can equivalently be interpreted in terms of activating an optimal subset of all the available sensors deployed in the spatial area (the non-activated sensors remain dormant). This framework is typical of the measurement regime in modern sensor networks.

A grave difficulty in selecting an optimal subset of gauged sites from among a given set of candidate sites is the combinatorial nature of this optimization problem. As the cardinalities of those sets increase, the exhaustive search of the search space quickly becomes computationally intractable. This stimulated attempts to solve this problem in a more constructive manner. For problems with low or moderate

dimensionalities, Uciński and Patan (2007) set forth a branch-and-bound method, which most often drastically reduces the search space and produces an optimal integral solution. In turn, for large-scale observation networks, the existing approaches replace the original NP-hard combinatorial problem with its convex relaxation in the form of a convex programming problem. This paves the way for application of interior-point methods, see the works by Joshi and Boyd (2009) and Chepuri and Leus (2015), or polyhedral approximation methods, cf. Uciński and Patan (2007), Herzog et al. (2018) and Uciński (2020b).

A major drawback of convex relaxation is the necessity of transforming the optimal relaxed solution into an acceptable solution of the original combinatorial problem. This is by no mean trivial and, when done carelessly, may make the performance gap between both the solutions quite wide. In recent years, handling the problem with various sparsity-enforcing penalty terms have won popularity. Although as a result of their addition to the original design criteria the problem convexity is lost, this property can be easily retrieved by resorting to iterations of the majorization-minimization scheme, cf. Sun et al. (2017). Unfortunately, it is not clear how to control the impact of this penalty so as not to depart from the relaxed solution too much.

The main contribution of this work consists in establishing an alternative approach which explicitly controls the quality of the sparsified solution in terms of the original design criterion. It focuses on directly minimizing the sparsity-enforcing penalty within the set of relaxed solutions which are allowed to deteriorate the optimal relaxed solution in terms of the original design criterion by no more than an arbitrarily set threshold. The technique reduces to a sequence of nonlinearly constrained convex optimization problems which are solved using an extremely fast generalized simplicial decomposition. As a result, a relatively simple and efficient technique of postprocessing relaxed solutions is proposed.

2. D-OPTIMUM SENSOR LOCATION AND ITS CONVEX RELAXATION

Consider a spatiotemporal system whose scalar state is given by the solution y to a deterministic partial differential equation (PDE) accompanied by the appropriate boundary and initial conditions. The PDE is defined on a bounded spatial domain $\Omega \subset \mathbb{R}^d$ $(d \leq 3)$ with a boundary $\partial\Omega$ and a bounded time interval $T = (0, t_f]$. It is specified up to $\boldsymbol{\theta} \in \mathbb{R}^m$, a vector of unknown parameters which are to be estimated from noisy observations of the state. These observations are going to be made by n pointwise sensors at given time instants $t_1, \ldots, t_K \in T$.

The sensor locations are to be selected from among $N > n$ candidate sites $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N \in \bar{\Omega} := \Omega \cup \partial\Omega$. Their measurements are modelled as

$$z_j = y(\boldsymbol{x}^{i_j}, t_k; \bar{\boldsymbol{\theta}}) + \varepsilon_{i_j,k} \tag{1}$$

for $j = 1, \ldots, n$ and $k = 1, \ldots, K$, where $y(\boldsymbol{x}, t; \boldsymbol{\theta})$ stands for the state at a spatial point $\boldsymbol{x} \in \bar{\Omega}$ and a time instant $t \in \bar{T} := [0, t_f]$, evaluated for a given parameter $\boldsymbol{\theta}$. Here $\bar{\boldsymbol{\theta}}$ signifies the vector of 'true' values of the unknown parameters, $i_1, \ldots, i_n \in \{1, \ldots, N\}$ are the indices of the gauged sites, and the $\varepsilon_{i_j,k}$'s are independent normally-distributed random errors with zero mean and constant variance σ^2.

Depending on whether or not prior background information about $\bar{\boldsymbol{\theta}}$ is accessible, Bayesian (e.g., maximum a posteriori estimation) or frequentionist (maximum-likelihood estimation, possibly combined with the Tikhonov-Phillips regularization) approaches can be used to produce an estimate $\hat{\boldsymbol{\theta}}$ of $\bar{\boldsymbol{\theta}}$. Its accuracy is characterized by the Fisher information matrix (FIM), cf. Atkinson et al. (2007)

$$\boldsymbol{M}(\boldsymbol{v}) = \boldsymbol{M}_0 + \sum_{i=1}^{N} v_i \boldsymbol{M}_i, \tag{2}$$

in which

$$\boldsymbol{M}_i = \frac{1}{\sigma^2} \sum_{k=1}^{K} \boldsymbol{g}(\boldsymbol{x}^i, t_k) \boldsymbol{g}^\top(\boldsymbol{x}^i, t_k), \quad i = 1, \ldots, n, \tag{3}$$

with $\boldsymbol{g}(\boldsymbol{x}, t) = \nabla_{\boldsymbol{\theta}}\, y(\boldsymbol{x}, t; \boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}$ stands for the sensitivity vector evaluated at a prior estimate $\boldsymbol{\theta}^0$ of the 'true' vector $\bar{\boldsymbol{\theta}}$. Furthermore, $\boldsymbol{v} = (v_1, \ldots, v_N)$ and v_i is the binary indicator variable equal to 1 or 0 depending on whether or not a sensor resides at site $\boldsymbol{x}^i$.

In the Bayesian setting, the inverse of $\boldsymbol{M}_0$ is the known covariance matrix of the prior density of $\boldsymbol{\theta}$ and then $\boldsymbol{M}^{-1}(\boldsymbol{v})$ approximates the covariance matrix of the posterior density of $\boldsymbol{\theta}$. In the frequentionist setting when no background information is available, $\boldsymbol{M}_0$ is set as zero, and then $\boldsymbol{M}^{-1}(\boldsymbol{v})$ approximates the covariance matrix of the maximum-likelihood estimator of $\bar{\boldsymbol{\theta}}$.

The 'goodness' of a sensor location represented as $\boldsymbol{v}$ is quantified by a design criterion related to the confidence ellipsoid, i.e., a highest probability regions for the parameters. The most common option is the D-optimality criterion

$$\Phi_{\mathrm{D}}(\boldsymbol{M}) = \det{}^{1/m}(\boldsymbol{M}), \tag{4}$$

maximization of which is equivalent to minimizing the volume of the confidence ellipsoid. Thus the spread of the estimates of the unknown parameters around their mean values is minimized (and on some not particularly restricted assumptions, these means are at least approximately equal to the 'true' values of the unknown parameters).

Consequently, we can formulate the following problem:

Problem 1. Find $\boldsymbol{v}^\star_{\mathrm{bin}} \in \mathcal{V}_{\mathrm{bin}} := \{\boldsymbol{v} \in \{0,1\}^N : \boldsymbol{1}^\top \boldsymbol{v} = n\}$ to maximize $\Phi_{\mathrm{D}}(\boldsymbol{M}(\boldsymbol{v}))$ Each minimizer $\boldsymbol{v}^\star_{\mathrm{bin}}$ is called a D-optimum *exact* design.

Unfortunately, the search for $\boldsymbol{v}^\star_{\mathrm{bin}}$ through evaluating $\Phi_{\mathrm{D}}(\boldsymbol{M}(\boldsymbol{v}))$ for each of $\binom{N}{n}$ possible choices of gauged sites quickly becomes computationally intractable with an increase in N. Therefore, following the customary procedure adopted in OED, we relax the nonconvex 0–1 constraints on the design variables, thereby allowing them to take any real values in the interval $[0, 1]$. Thus we get the following much more convenient formulation:

Problem 2. Find $\boldsymbol{v}^\star_{\mathrm{D}} \in \mathcal{V} := \{\boldsymbol{v} \in [0,1]^N : \boldsymbol{1}^\top \boldsymbol{v} = n\}$ to maximize $\Phi_{\mathrm{D}}(\boldsymbol{M}(\boldsymbol{v}))$ Each minimizer $\boldsymbol{v}^\star_{\mathrm{D}}$ is called a D-optimum *relaxed* design.

3. SPARSITY-PROMOTING POSTPROCESSING OF RELAXED DESIGNS

Obviously, the relaxed design $\boldsymbol{v}_{\mathrm{D}}^{\star}$ may have many fractional components. On the one hand, there is an acute need to convert it to a design desirably with binary weights. Mathematically, the transformed design ought to be characterized by minimal support, i.e., a minimal number of nonzero components. On the other hand, however, the loss in the attained extreme value of the D-optimality criterion should be as small as possible.

For any design $\boldsymbol{v} \in \mathcal{V}$ define its D-efficiency

$$\mathcal{E}_{\mathrm{D}}(\boldsymbol{v}) = \frac{\det^{1/m}(\boldsymbol{M}(\boldsymbol{v}))}{\det^{1/m}(\boldsymbol{M}(\boldsymbol{v}_{\mathrm{D}}^{\star}))}, \tag{5}$$

cf. Atkinson et al. (2007). It quantifies the degree of D-optimality of $\boldsymbol{v}$ with respect to $\boldsymbol{v}_{\mathrm{D}}^{\star}$. Note that $0 \leq \mathcal{E}_{\mathrm{D}}(\boldsymbol{v}) \leq 1 = \mathcal{E}_{\mathrm{D}}(\boldsymbol{v}_{\mathrm{D}}^{\star})$.

Let us also denote by $\|\boldsymbol{v}\|_0$ the number of nonzero components of the vector $\boldsymbol{v}$, i.e., $\|\boldsymbol{v}\|_0 = |\mathrm{supp}(\boldsymbol{v})|$, $\mathrm{supp}(\boldsymbol{v}) = \big\{j \in \{1,\ldots,N\} : v_j \neq 0\big\}$,

The following formulation can address the above dilemma:

Problem 3. Find $\boldsymbol{v}_{\mathrm{D},0}^{\star} \in \mathcal{V} := \big\{\boldsymbol{v} \in [0,1]^N : \boldsymbol{1}^{\top}\boldsymbol{v} = n\big\}$ to minimize $\|\boldsymbol{v}\|_0$ subject to

$$\mathcal{E}_{\mathrm{D}}(\boldsymbol{v}) \geq \eta, \tag{6}$$

where η signifies a given minimal acceptable value of D-efficiency. Each minimizer $\boldsymbol{v}_{\mathrm{D},0}^{\star}$ is called a minimum-support *relaxed* design with guaranteed D-efficiency η.

The constraint (6) endows Problem 3 with the following meaningful interpretation: Having obtained a D-optimum design $\boldsymbol{v}_{\mathrm{D}}^{\star}$, we sacrifice some (possibly high) degree of D-optimality, which is controlled by η, for minimization of the support of the relaxed design.

In the literature, $\|\cdot\|_0$ is termed the ℓ_0-norm, although it is not a norm, as it does not satisfies the condition of absolute homogeneity. Its direct minimization is extremely hard, which is why we replace the ℓ_0-'norm' by the ℓ_q-'norm'

$$\|\boldsymbol{w}\|_q = \Big(\sum_{j=1}^{r} |w_j|^q\Big)^{\frac{1}{q}}, \tag{7}$$

which is justified by the property $\|\boldsymbol{w}\|_0 = \lim_{q \downarrow 0} \|\boldsymbol{w}\|_q$ valid for each fixed $\boldsymbol{w}$. Observe that given $0 < q < 1$, $\|\cdot\|_q$ it is not a norm, either, as it does not satisfy the triangle inequality.

Problem 3 is then replaced by its counterpart for a fixed $q \in (0,1)$:

Problem 4. Find a vector $\boldsymbol{v}_{\mathrm{D},q}^{\star} \in \mathbb{R}^N$ minimizing

$$J(\boldsymbol{v}) = \|\boldsymbol{v}\|_q^q = \sum_{j=1}^{N} w_j^q \tag{8}$$

over the feasible set $\mathcal{W} := \{\boldsymbol{v} \in \mathcal{V} : \mathcal{E}_{\mathrm{D}}(\boldsymbol{v}) \geq \eta\}$.

3.1 Majorization-minimization algorithm

At first sight the concavity of the design criterion J seems to make the attendant optimization problems prohibitively difficult. However, on close examination the majorization-minimization (MM) algorithm, cf. Sun et al. (2017), turns out a simple remedy to this problem.

For J the MM algorithm is going to generate a sequence of feasible vectors $\big\{\boldsymbol{v}^{(\kappa)}\big\}_{\kappa=0}^{\infty}$ starting from an arbitrary feasible initial point $\boldsymbol{v}^{(0)}$ by minimizing in each iteration a surrogate function $\boldsymbol{v} \mapsto \Psi(\boldsymbol{v}|\boldsymbol{v}^{(\kappa)})$. This function should be a convex tangent majorant of $J(\boldsymbol{v})$ at $\boldsymbol{v}^{(\kappa)}$. Observe that here, in view of the concavity of J, we have

$$\begin{aligned} J(\boldsymbol{v}) &\leq J(\boldsymbol{v}^{(\kappa)}) + (\boldsymbol{v} - \boldsymbol{v}^{(\kappa)})^{\top}\nabla J(\boldsymbol{v}^{(\kappa)}) \\ &:= \Psi(\boldsymbol{v}|\boldsymbol{v}^{(\kappa)}), \quad \forall \boldsymbol{v} \in \mathcal{W}, \end{aligned} \tag{9}$$

and its right-hand side satisfies all the requirements for the surrogate function in question.

The consecutive iterates of the MM algorithm are then defined as

$$\boldsymbol{v}^{(\kappa+1)} = \arg\min_{\boldsymbol{v} \in \mathcal{W}} \Psi(\boldsymbol{v}|\boldsymbol{v}^{(\kappa)}). \tag{10}$$

A characteristic feature of the MM algorithm is that the sequence $\big\{J(\boldsymbol{v}^{(\kappa)})\big\}$ is strictly decreasing until a minimizer is attained. What is more, any limit point $\boldsymbol{v}^{\star}$ of $\big\{\boldsymbol{v}^{(\kappa)}\big\}$ is a stationary point of J. Note, however, that the function J may have multiple local minima and only one of them can be attained.

4. GENERALIZED SIMPLICIAL DECOMPOSTION TO MINIMIZE THE SURROGATE FUNCTION

The linearity of the surrogate objective function $\Psi(\,\cdot\,|\boldsymbol{v}^{(\kappa)})$, the polyhedral form of the set $\mathcal{V}$ and the convexity of the constraint (6) make the optimization problem (10) ideal for the use of generalized simplicial decomposition (GSD) to quickly solve it and drastically reduce the problem dimensionality, cf. Bertsekas and Yu (2011) and applications in Uciński (2020a) and Uciński (2022). The essence of the dimensionality reduction involved by GSD is that at each iteration the polyhedral set $\mathcal{V}$ in $\mathbb{R}^N$ is approximated with the convex hull of an ever expanding set $\mathcal{V}^{(\tau)}$ that consists of extreme points of $\mathcal{V}$. The method alternates between minimization of $\Psi(\boldsymbol{v}|\boldsymbol{v}^{(\kappa)})$ over $\mathrm{conv}(\mathcal{V}^{(\tau)})$, subject to the additional side constraint (6), and finding a new extreme point $\boldsymbol{v}_{\mathrm{xtrm}}^{(\tau)} \in \mathcal{V}$ to be added to $\mathcal{V}^{(\tau)}$. (These tasks are called the restricted master problem, or RMP, and the column generation problem, or CGP, respectively.) The polyhedron $\mathrm{conv}(\mathcal{V}^{(\tau)})$ can be treated as a subset of $\mathbb{R}^{\tau+2}$ and for a typical run of GSD we have $\tau \ll N$. This is where a substantial dimensionality reduction emerges.

The low-dimensional RMP can be solved by any solver for constrained nonlinear optimization which returns, as a by-product, the values of the Lagrange multipliers (this is usually the case when fast Newton-like methods, such as SQP, are used).

5. NUMERICAL RESULTS

Consider the setting of the moving source identification investigated in the paper by Uciński (2022). There are seven unknown parameters characterizing the diffusion coefficient and the moving source. The parameters are going to be estimated using $n = 150$ sensors selected from a total of $N = 780$ sensor network nodes residing

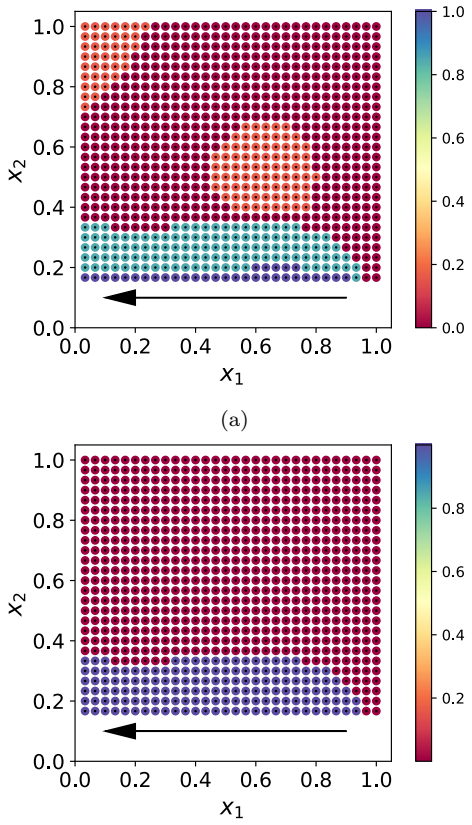


Fig. 1. Optimum sensor configurations: relaxed D-optimum design with some nonnegligible weights in the centre and top-left subregions (a), sparse solution with a guaranteed D-efficiency of $\eta = 0.95$ (b). Black points denote candidate sites and the colours of the discs around them represent the weight values. The arrow represents the movement of the pollution source.

in a unit square. Figure 1 shows the original relaxed D-optimum design weights and the same weights post-processed with the proposed algorithm for the minimal acceptable D-efficiency level set at 0.95 and $q = 0.2$. Only two iterations of the MM scheme were needed. Overall, the postprocessing took no more than 1 second on a mediocre PC computer and an implementation in Python.

REFERENCES

Alexanderian, A. (2021). Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: A review. *Inverse Problems*, 37(4), 043001.

Alexanderian, A., Petra, N., Stadler, G., and Ghattas, O. (2014). A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification. *SIAM Journal on Scientific Computing*, 36(5), A2122–A2148. doi: 10.1137/130933381.

Atkinson, A.C., Donev, A.N., and Tobias, R.D. (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press, Oxford.

Bertsekas, D. and Yu, H. (2011). A unifying polyhedral approximation framework for convex optimization. *SIAM Journal on Optimization*, 21(1), 333–360. doi: 10.1137/090772204.

Chepuri, S.P. and Leus, G. (2015). Sparsity-promoting sensor selection for non-linear measurement models. *IEEE Transactions on Signal Processing*, 63(3), 684–698.

Gejadze, I.Y. and Shutyaev, V. (2012). On computation of the design function gradient for the sensor-location problem in variational data assimilation. *SIAM Journal on Scientific Computing*, 34(2), B127–B147. doi: 10.1137/110825121.

Haber, E., Horesh, L., and Tenorio, L. (2010). Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems. *Inverse Problems*, 26(2), 025002.

Herzog, R., Riedel, I., and Uciński, D. (2018). Optimal sensor placement for joint parameter and state estimation problems in large-scale dynamical systems with applications to thermo-mechanics. *Optimization and Engineering*, 19(3), 591–627. doi: https://doi.org/10.1007/s11081-018-9391-8.

Joshi, S. and Boyd, S. (2009). Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2), 451–462.

Patan, M. (2012). *Optimal Sensor Networks Scheduling in Identification of Distributed Parameter Systems*. Springer-Verlag, Berlin.

Rafajłowicz, E. (2022). *Optimal Input Signals For Parameter Estimation: In Linear Systems with Spatio-Temporal Dynamics*. De Gruyter, Berlin.

Sun, Y., Babu, P., and Palomar, D.P. (2017). Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3), 794–816.

Uciński, D. (2005). *Optimal Measurement Methods for Distributed-Parameter System Identification*. CRC Press, Boca Raton, FL.

Uciński, D. (2020a). Construction of constrained experimental designs on finite spaces for a modified E_k-optimality criterion. *International Journal of Applied Mathematics and Computer Science*, 30(4), 659–677. doi:https://doi.org/10.34768/amcs-2020-0049.

Uciński, D. (2020b). D-optimal sensor selection in the presence of correlated measurement noise. *Measurement*, 164, 107873. doi: https://doi.org/10.1016/j.measurement.2020.107873.

Uciński, D. (2022). E-optimum sensor selection for estimation of subsets of parameters. *Measurement*, 187, 110286. doi: https://doi.org/10.1016/j.measurement.2021.110286.

Uciński, D. and Patan, M. (2007). D-optimal design of a monitoring network for parameter estimation of distributed systems. *Journal of Global Optimization*, 39(2), 291–322.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Structure of Feedback-Loops through Positive Real Odd Functions

Izchak Lewkowicz *

* *School of Electrical and Computer Engineering, Ben-Gurion University, Beer-Sheba, Israel (e-mail: izchak@bgu.ac.il)*

The family of (matrix-valued) Positive real Odd, $\mathcal{PO}$, functions $F(s)$ may be described as,

$$\mathcal{PO} := \{F(s) \ : \ F(s) = -\left(F(-s^*)\right)^* \quad \forall s \in \mathbb{C} \ \}.$$

Resorting to the framework of the Quadratic Matrix Inequality one can explicitly describe $\mathcal{PO}$ functions as,

$$\mathcal{PO} = \left\{ F(s) : \begin{bmatrix} F(s) \\ I_m \end{bmatrix}^* \begin{bmatrix} 0 & I_m \\ I_m & 0 \end{bmatrix} \begin{bmatrix} F(s) \\ I_m \end{bmatrix} \begin{matrix} \preccurlyeq 0 & \forall s \in \mathbb{C}_L \\ = 0 & \forall s \in i\mathbb{R} \\ \succcurlyeq 0 & \forall s \in \mathbb{C}_R \end{matrix} \right\}. \quad (1)$$

The subset of rational functions within this family (a.k.a. Lossless or Foster) corresponds to electrical circuits with reactive elements (i.e. $L - C$) and no resistance, see e.g. Figures 1, 2 and 5, below.

In Figure 1 the elements sC and sL on the left-hand side, are substituted, on the right, by an admittance network Y_F and an impedance network Z_G, respectively. In contrast, in Figure 2 the elements sC and sL on the left-hand side are substituted, on the right, by an admittance network Y_G and an impedance network Z_F, respectively.

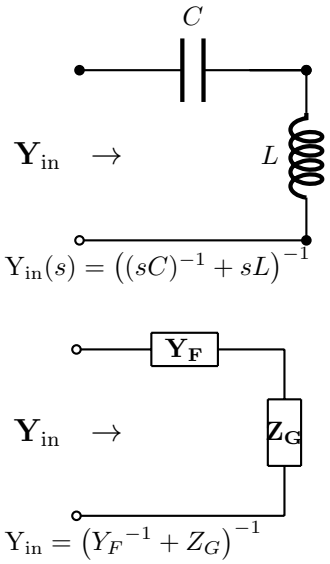


Fig. 1. Driving point admittance of a series circuit of degree 2

For instance, the elements sC and sL, on the left-hand side of Figure 1 are substituted, on the right-hand side, by admittance network Y_F and impedance network Z_G, respectively. In contrast, in Figure 2 sC and sL, on the left-hand side, are substituted, on the right-hand side, by the admittance network Y_G and impedance network Z_F, respectively.

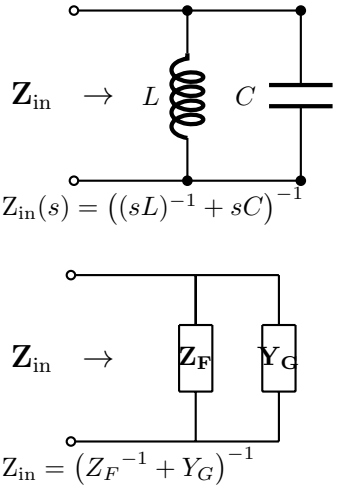


Fig. 2. Driving point impedance of a parallel circuit of degree 2

It is now appropriate to introduce the notation,

$$\phi(X, Y) := \left(X^{-1} + Y\right)^{-1}, \quad (2)$$

where X, Y are square variables so that X and $(X^{-1}+Y)$ are non-singular.

As a first application, one can write,

Figure 1	Figure 2
$Y_{in}(s)=\phi(sC \ , \ sL)$	$Z_{in}(s)=\phi(sL \ , \ sC)$
$Y_{in}(s)=\phi(Y_F \ , \ Z_G)$	$Z_{in}(s)=\phi(Z_F, \ Y_G).$

The rest of this section is devoted to showing that the role of the function ϕ in Eq. (2), goes beyond an exercise in circuits.

Following the spirit of the description of $\mathcal{PO}$ functions in Eq. (1), ϕ is a $\mathcal{PO}$ function in the sense that,

$$(X+X^*) \ \text{ and } \ (Y+Y^*) \begin{matrix} \preccurlyeq 0 \\ = 0 \\ \succcurlyeq 0 \end{matrix} \implies \Phi+\Phi^* \begin{matrix} \preccurlyeq 0 \\ = 0 \\ \succcurlyeq 0 \end{matrix} \quad (3)$$

Note also that whenever in addition Y and $(Y^{-1}+X)$ are non-singular then,

$$\phi(Y, X) := \left(Y^{-1} + X\right)^{-1} = \phi(X^{-1}, Y^{-1}). \quad (4)$$

Thus, one can conclude that ϕ is a $\mathcal{PO}$ function of two non-commuting variables.

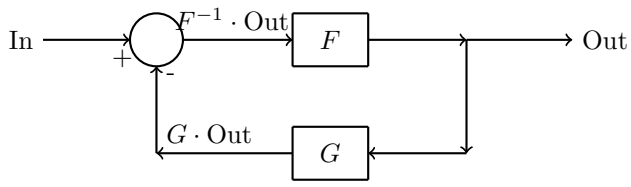


Fig. 3. Out $= \left(F^{-1}+G\right)^{-1}\cdot \text{In} = \phi(F,\ G)\cdot \text{In}$

In analogy to Figures 1 and 2, one may view Figure 4 as a complement to Figure 3.

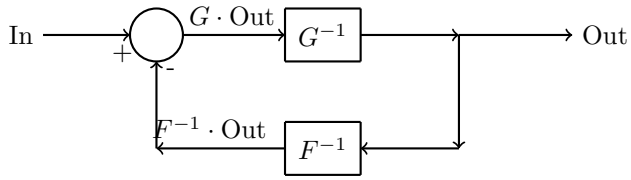


Fig. 4. Out $= \left(F^{-1}+G\right)^{-1}\cdot \text{In} = \phi(G^{-1},\ F^{-1})\cdot \text{In}$

Let F and G are $m\times m$-valued rational functions. Assuming in Figure 3 that F and $(F^{-1}+G)$ are invertible, or assuming in Figure 4 that in addition G is invertible, one has that in both cases,

$$\begin{aligned}\text{In}-G\cdot\text{Out}=F^{-1}\cdot\text{Out} \quad &\Longrightarrow \quad \text{In}=\left(F^{-1}+G\right)\cdot\text{Out}\\ &\Longrightarrow \quad \text{Out}=\phi(F,G)\cdot\text{In}.\end{aligned}$$

In each of the Figures 1 and 2 the circuit was comprised of a pair reactive element, i.e. a rational positive real odd function of degree two. The circuit in Figure 5 is comprised of four reactive elements.

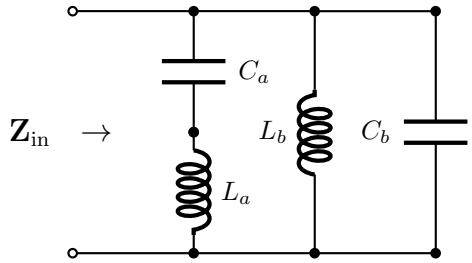


Fig. 5. $\text{Z}_{\text{in}}(s) = \left(\left((sC_a)^{-1}+sL_a\right)^{-1}+(sL_b)^{-1}+sC_b\right)^{-1}$.

The driving point impedance in Figure 5 can also be written as,

$$\begin{aligned}\text{Z}_{\text{in}}(s) &= \Big(\overbrace{\left((sC_a)^{-1}+sL_a\right)^{-1}}^{\phi(sC_a\ ,\ sL_a)}+\overbrace{{sL_b}^{-1}+sC_b}^{(\phi(sL_b\ ,\ sC_b))^{-1}}\Big)^{-1} \qquad (5)\\ &= \phi\left(\ \phi(sL_b\ ,\ sC_b),\ \phi(sC_a\ ,\ sL_a)\ \right).\end{aligned}$$

As before, in analogy to the $L-C$ circuit in Figure 5, we have the multiple feedback loops in Figure 6.

As before one can re-write the input-output relation in Figure 6 as,

$$\begin{aligned}\text{Out} &= \left(\left({F_b}^{-1}+G_b\right)^{-1}+{F_a}^{-1}+G_a\right)^{-1}\text{In}\\ &= \phi(\ \phi(F_a\ ,\ G_a)\ ,\ \phi(F_b\ ,\ G_b)\ )\text{In}.\end{aligned} \qquad (6)$$

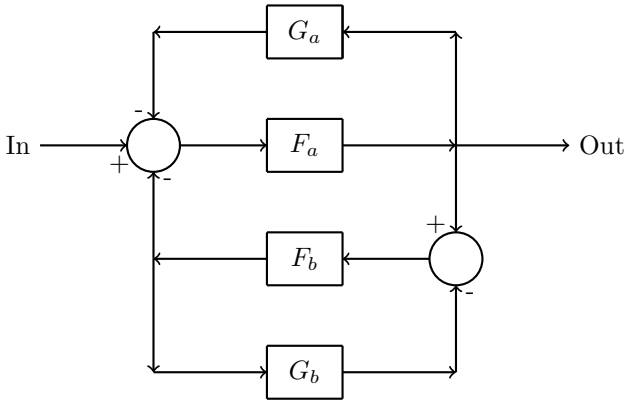


Fig. 6. Out $= \left(\left({F_b}^{-1}+G_b\right)^{-1}+{F_a}^{-1}+G_a\right)^{-1}$ In

Now, comparing Eq. (5) with Eq. (6), one can formally identify the elements sC_a, sL_a, sC_b, sL_b, in Figure 5 with the blocks $F_a(s)$, $G_a(s)$, $F_c(s)$, $F_d(s)$ in Figure 6.

This calls for adapting one of the classical construction schemes of $R-L-C$ circuits, e.g. Brune, Bott-Duffin, Darlington, Foster, Cauer, etc., to introducing a design tool for networks of feedback-loops, more elaborate than that in Figure 6 (and as mentioned, the building blocks need not be positive real).

A word of caution: The passage from one-port circuit design to that of feedback-loops networks can not be straightforward: Typically blocks like $F_a(s)$, $G_a(s)$, $F_c(s)$, $F_d(s)$ are *non-commutative*. Hence, one needs to formally extend the notion of $\mathcal{PO}$ functions to real rational functions of say k *non-commuting* variables. See e.g. ϕ in Eqs. (2) (3)

As a hint to the potential of the propsed approach, the feedback loop network from Figure 6 is extended in Figure 7, to having two inputs and two outputs.

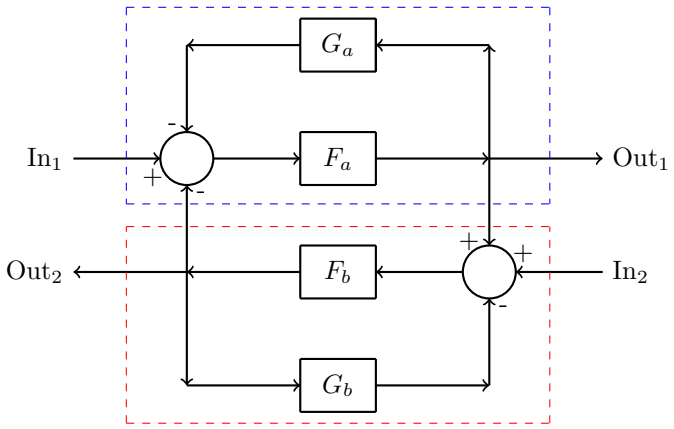


Fig. 7. Multiple Feedback Loops

$$\begin{aligned}\begin{bmatrix}\text{Out}_1\\ \text{Out}_2\end{bmatrix} &= \begin{bmatrix}\left(F_c+G_c^{-1}\right)^{-1} & -\left(F_c+G_c^{-1}\right)^{-1}G_c^{-1}\\ G_c^{-1}\left(F_c+G_c^{-1}\right)^{-1} & \left(F_c^{-1}+G_c\right)^{-1}\end{bmatrix}\begin{bmatrix}\text{In}_1\\ \text{In}_2\end{bmatrix}\\ &= \begin{bmatrix}\left(F_c+G_c^{-1}\right)^{-1} & -F_c^{-1}\left(F_c^{-1}+G_c\right)^{-1}\\ \left(F_c^{-1}+G_c\right)^{-1} & F_c^{-1}\left(F_c^{-1}+G_c\right)^{-1}\end{bmatrix}\begin{bmatrix}\text{In}_1\\ \text{In}_2\end{bmatrix}\end{aligned} \qquad (7)$$

where

$$F_c := F_a^{-1} + G_a \qquad G_c := F_b^{-1} + G_b \ .$$

REFERENCES

[1] I. Lewkowicz, "Passive Linear Continuous-time Systems- Characterization through Structure", *Systems and Control Letters*, Vol. 147, No. 104819, 2021.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Realizations of Passive Systems are Inter-related

Izchak Lewkowicz

School of Electrical and Computer Engineering
Ben-Gurion University, Beer-Sheba, Israel
E-mail: izchak@bgu.ac.il

Passive Rational Functions

Passivity is a physical property of dynamical systems. Recently, it was shown that, at least in the linear time-invariant framework, it is equivalent to an explicit mathematical structure. See **?**, **?**.

As it is well known, physically the set of $m \times m$-valued positive real rational functions represents linear time-invariant passive systems.

In fact, there are four variants. To describe them we shall adopt the following notation. $\mathbb{C}_R$ will be the open right half of the complex plane, and $\overline{\mathbb{D}}^c = \{c \in \mathbb{C} \ : \ |c| > 1 \ \}$, i.e. the exterior of the closed unit disk.
Consider $m \times m$-valued real rational functions $F(z)$ (m is a parameter).

$\mathcal{P}$ Positive-Real (continuous-time) $\mathrm{Re}\,(x^*F(s)x) \geq 0 \ \ \forall x \in \mathbb{C}^m \ \ \forall s \in \mathbb{C}_R$

$\mathcal{B}$ Bounded-Real, (continuous-time) $1 \geq \|F(s)\|_2 \ \ \forall s \in \mathbb{C}_R$.

$\mathcal{DP}$ Discrete-Time-Positive-Real $\mathrm{Re}\,(x^*F(s)x) \geq 0 \ \ \forall x \in \mathbb{C}^m \ \ \forall s \in \overline{\mathbb{D}}^c$

$\mathcal{DB}$ Discrete-Time-Bounded-Real $1 \geq \|F(s)\|_2 \ \ \forall s \in \overline{\mathbb{D}}^c$

The above four families, are common in Engineering circles, in particular $\mathcal{P}$ can be viewed as the Laplace transform of a continuous-time, stable, linear time-invariant system, described by a differential equation, while $\mathcal{DB}$ may be the Z-transform of a discrete-time, stable, linear time-invariant system, described by a difference equation.

As already mentioned, in **?**, **?** the above two better known families were characterized through their structure.

Theorem 1
Consider two families of functions.

$\mathcal{P}$ *This set is a maximal matrix-convex cone of matrix-valued real rational functions, closed under inversion, which are analytic in* $\mathbb{C}_R$.

Conversely, a maximal matrix-convex cone of matrix-valued rational functions, closed under inversion, which are analytic in $\mathbb{C}_R$ *and containing the zero degree function* $F_o(s) \equiv I$*, is the set* $\mathcal{P}$.

$\mathcal{DB}$ *Let* $\mathcal{F}$ *be a family of square matrix-valued (of various dimensions) real rational functions* $F(s)$*. For all* s *outside the closed unit disk, each* $F(s)$ *is analytic.*

If as a family, $\mathcal{F}$ *is matrix-convex, and a maximal set closed under products of its elements (whenever dimensions are suitable), this is the set* $\mathcal{DB}$ *of Discrete-time Bounded real rational functions.*

The converse is true as well.

By definition each of the pairs $\mathcal{P}$, $\mathcal{B}$ along with $\mathcal{DP}$, $\mathcal{DB}$, shares the same domain. In contrast, each of the pairs $\mathcal{P}$, $\mathcal{DP}$ along with $\mathcal{B}$, $\mathcal{DB}$ shares the same range. To be more specific, we recall the following: Assume that the Cayley transform of $F(s)$ is, almost everywhere, well defined, i.e.

$$\mathcal{C}(F) = (I_m - F)(I_m + F)^{-1} \qquad \det\left(F(s) + I_m\right) \not\equiv 0. \quad (1)$$

Then above four variants of passive functions satisfy the following relations,

$$\mathcal{B} = \mathcal{C}\left(\mathcal{P}\right) \qquad \mathcal{DB} = \mathcal{C}\left(\mathcal{DP}\right)$$

$$F(s) \in \mathcal{P} \iff F\left(\tfrac{1+s}{1-s}\right) \in \mathcal{DP} \quad (2)$$

$$F(s) \in \mathcal{B} \iff F\left(\tfrac{1+s}{1-s}\right) \in \mathcal{DB}.$$

Namely the sets $\mathcal{P}$, $\mathcal{B}$, $\mathcal{DP}$ and $\mathcal{DB}$ are inter-related as *rational functions.*

Here we proceed in the direction of exploring structure of *families* passive systems, and now focus on state-space *realizations*. Specifically, we show that realizations of the above four families, are inter-related through *Linear Fractional Transformation*. Specifically, starting with any of the families ($\mathcal{P}$, $\mathcal{B}$, $\mathcal{DP}$ or $\mathcal{DB}$) all other three can be obtained.

Here are the details. Recall that with an arbitrary $m \times m$-valued rational function $F(s)$, with no pole at infinity, i.e.

$$\exists \lim_{s \longrightarrow \infty} F(s),$$

one can associate a corresponding $(n+m) \times (n+m)$ state-space realization array, R_F i.e.

$$F(s) = C(sI_n - A)^{-1}B + D \qquad R_F = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]. \quad (3)$$

The realization R_F in Eq. (3) is called *minimal*, if n is the McMillan degree of $F(s)$.

We can now state the main result of this presentation.

Theorem 2
For $l = \mathcal{P}$, $\mathcal{B}$, $\mathcal{DP}$ and $\mathcal{DB}$, let $\mathcal{F}_l$ be families of rational functions. Let R_{F_l} be a corresponding state space realization (assuming there is no pole at infinity). Then (whenever inverses exit) the following is true.

$$(\mathcal{P}) \quad R_{F_\mathcal{P}} = \begin{bmatrix} A_{\mathcal{DB}}-I_n & B_{\mathcal{DB}} \\ -C_{\mathcal{DB}} & -D_{\mathcal{DB}}+I_m \end{bmatrix} \begin{bmatrix} A_{\mathcal{DB}}+I_n & B_{\mathcal{DB}} \\ C_{\mathcal{DB}} & D_{\mathcal{DB}}+I_m \end{bmatrix}^{-1}$$

$$(\mathcal{B}) \quad R_{F_\mathcal{B}} = \begin{bmatrix} A_\mathcal{P}-B_\mathcal{P}(I_m+D_\mathcal{P})^{-1}C_\mathcal{P} & -\sqrt{2}B_\mathcal{P}(I_m+D_\mathcal{P})^{-1} \\ \sqrt{2}(I_m+D_\mathcal{P})^{-1}C & (I_m-D_\mathcal{P})(I_m+D_\mathcal{P})^{-1} \end{bmatrix}$$

$$(\mathcal{DP}) \quad R_{F_\mathcal{DP}} = \begin{bmatrix} A_\mathcal{B}+I_n & B_\mathcal{B} \\ C_\mathcal{B} & D_\mathcal{B}-I_m \end{bmatrix} \begin{bmatrix} A_\mathcal{B}-I_n & B_\mathcal{B} \\ -C_\mathcal{B} & -D_\mathcal{B}-I_m \end{bmatrix}^{-1}$$

$$(\mathcal{DB}) \quad R_{F_\mathcal{DB}} = \begin{bmatrix} A_\mathcal{DP}-B_\mathcal{DP}(I_m+D_\mathcal{DP})^{-1}C_\mathcal{DP} & -\sqrt{2}B_\mathcal{DP}(I_m+D_\mathcal{DP})^{-1} \\ \sqrt{2}(I_m+D_\mathcal{DP})^{-1}C & (I_m-D_\mathcal{DP})(I_m+D_\mathcal{DP})^{-1} \end{bmatrix}.$$

Roughly speaking, one can conclude that K-Y-P type results for $\mathcal{P}$, $\mathcal{B}$, $\mathcal{DP}$ or $\mathcal{DB}$, are all equivalent, up to linear fractional transformation.

Although derived independently, one can relate Eq. (2) to Theorem 2.

Remarks
a. Having the "matrix" R_F non-singular, and the realization array R_F minimal, are two independent properties.

b. Recall in the realization of a Cayley transform of a rational function: Let $F(s)$ be a square matrix-valued rational function along with its realization as in Eq. (3). Then a realization of $\mathcal{C}(F)$ is given by,

$$R_{\mathcal{C}(F)} = \left[\begin{array}{c|c} A-B(I_m+D)^{-1}C & \mp\sqrt{2}B(I_m+D)^{-1} \\ \hline \pm\sqrt{2}(I_m+D)^{-1}C & (I_m-D)(I_m+D)^{-1} \end{array}\right]. \quad (4)$$

This conforms with items $(\mathcal{B})$ and $(\mathcal{DB})$ in Theoremi 2.

c. From Eq. (3) one has that $F(s) \in \mathcal{C}(\mathcal{P}) \iff F\left(\frac{1+s}{1-s}\right) \in \mathcal{DB}$.
Using Theorem **??**, let $R_{F_\mathcal{P}} = \left[\begin{array}{c|c} A_\mathcal{P} & B_\mathcal{P} \\ \hline C_\mathcal{P} & D_\mathcal{P} \end{array}\right]$ and $R_{F_\mathcal{DB}} = \left[\begin{array}{c|c} A_\mathcal{DB} & B_\mathcal{DB} \\ \hline C_\mathcal{DB} & D_\mathcal{DB} \end{array}\right]$ be realizations of $\mathcal{P}$ and $\mathcal{DB}$ functions, respectively. Then whenever inverses exit, $R_{F_\mathcal{P}} = -\mathcal{C}\left(R_{(F_\mathcal{DB})^{-1}}\right)$.

References:

I. Lewkowicz, "Passive Linear Continuous-time Systems-Characterization through Structure", *Systems and Control Letters*, Vol. 147, No. 104819, 2021.

I. Lewkowicz, "Passive Linear Discrete-time Systems - Characterization through Structure", *Linear Algebra and its Applications*, No. 15643, 2021.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Nonlinear Mechanical Network-based Models for Force-controlling Devices ⋆

Yuan Li * **Xiaofu Liu** ** **Jason Zheng Jiang** *
Branislav Titurus * **Simon Neild** *

* *School of Civil, Aerospace and Mechanical Engineering, University of Bristol, Bristol, BS8 1TR, UK (e-mail: yl14470@bristol.ac.uk, z.jiang@bristol.ac.uk, brano.titurus@bristol.ac.uk, simon.neild@bristol.ac.uk).*
** *College of Engineering, China Agricultural University, Beijing, China (e-mail: lxf@cau.edu.cn)*

Abstract: In recent years, network synthesis theory has been successfully applied to vibration absorber design, to identify optimum mechanical networks providing performance improvements. These identified mechanical networks consist of ideal linear modelling elements, such as springs, dampers and inerters. For real-life applications, the essential next step is to transfer these linear mechanical networks into physical absorber designs. There are two major challenges for this step: firstly, in order to achieve practical physical realisations, multidomain physical components (mechanical, hydraulic, pneumatic and electrical) need to be considered; and secondly, nonlinearities and other parasitic properties of physical components must be taken into consideration or potentially be made full use of. To this end, this paper, using a nonlinear mechanical network-based model for a bespoke mechanical-hydraulic device, demonstrates the feasibility of resolving both challenges.

Keywords: Vibration absorber design; nonlinearity; force control; mechanical network; physical models

1. INTRODUCTION

For vibration absorber design, different from the traditional approach focusing on modifying the specific designs, network synthesis theory has been applied in recent years, allowing a wide range of absorber networks to be explored systematically (Smith, 2002; Chen and Smith, 2009; Yamamoto and Smith, 2015; Zhang et al., 2017b, 2019a; Hughes, 2020). These networks consist of ideal linear modelling properties, such as stiffness, damping and inertance (Smith, 2002) and have demonstrated their advantages for various engineering structures (Smith and Wang, 2004; Jiang et al., 2012; Zhang et al., 2017a, 2019b; Lewis et al., 2019). To achieve the real-life benefits, the next step is to transfer these networks into physical absorber design.

Not only focusing on mechanical domain, the physical realisation needs to consider the components from multiple domains (mechanical, hydraulic, pneumatic and electrical) allowing numerous design possibilities and functionalities to be explored. There are challenges to be addressed for this step. The first challenge is to achieve the ideal mechanical network topology, which depends on how to choose the multidomain components and coupling mechanisms. The basic modelling elements from each domain have been summarised in the table of Fig. 1 and the physical component would need single or multiple basic elements to represent, such as a hydraulic orifice as a single resistance while a hydraulic tube as hydraulic inertance and resistance in parallel (Liu et al., 2018, 2019). In order to achieve the desirable force-velocity properties, suitable coupling mechanisms need to be considered, such as the piston-cylinder coupling (Liu et al., 2019), rubber-fluid coupling (Li et al., 2019) and electrical transducers (Wang and Chan, 2011). These mechanisms will determine how the non-mechanical components and implementation are mapped into equivalent mechanical components and topological connections.

Once the desirable network topology is realised using multidomain components, the next challenge will be how to realise the network modelling properties (stiffness, damping and inertance) and the difficulty lies in that the physical component nonlinearities and parasitic effects need to be characterised. The approach proposed by Liu et al. (2019) which develops a generalisable model for a physical device with key design parameters will be used to address this challenge. This generalisable model will not only help to develop an accurate dynamic model but also allow the key physical parameters to be optimum designed maximising the performance benefits.

To address the aforementioned two challenges, a nonlinear mechanical network-based model is needed, to transfer the multidomain components into an equivalent mechanical network and enable the component nonlinearity to be characterised or potentially made use of. In this work, a feasibility of resolving these two challenges will be

⋆ Research is supported by the Engineering and Physical Sciences Research Council (EPSRC) (Grant Reference: EP/P013546/1).

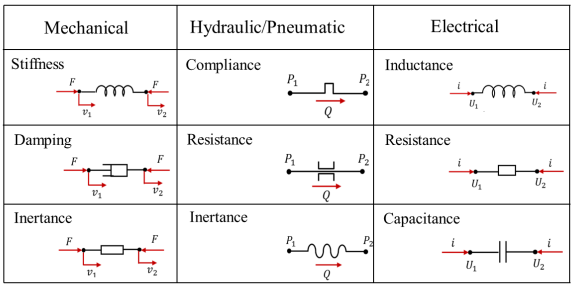


Fig. 1. A mechanical-hydraulic-electrical analogy (Schönfeld, 1954)

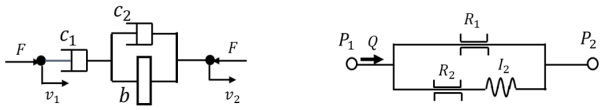
demonstrated using nonlinear mechanical network-based model for a bespoke mechanical-hydraulic device.

2. A FEASIBILITY STUDY USING A MECHANICAL-HYDRAULIC DEVICE

In this feasibility study, a mechanical-hydraulic device which could realise a specific linear mechanical network topology is first presented in Section 2.1. Based on this physical device, the component-level nonlinearites and parasitic effects are identified experimentally in Section 2.2. By integrating such component nonlinearities and parasitic effects into the mechanical network topology, the final nonlinear mechanical network-based model, where the properties are represented by physical design parameters, is developed in Section 2.3. Readers could refer to Liu et al. (2019) for more details.

2.1 Mutlidomain Realisation of A Linear Network

A linear mechanical network shown in Fig. 2(a) is assumed to be realised in this feasibility study.



(a) (b)

Fig. 2. (a) The linear mechanical network which needs to be realised; (b) The equivalent linear hydraulic network using the piston-cylinder coupling.

Considering a mechanical-hydraulic device and piston-cylinder coupling mechanism, the physical device force F and the relative terminal velocity Δv can be transferred to hydraulic pressure difference Δp and flow rate Q due to the piston movement, using the relationships:

$$F = A_P \Delta p, \tag{1}$$

$$\Delta v = \frac{Q}{A_P}, \tag{2}$$

where A_P is the piston area. Using these coupling equations, the mechanical through variable F is linked with the hydraulic cross variable Δp while the mechanical cross variable Δv with the hydraulic through one Q. Therefore, the equivalent hydraulic network is dual to the mechanical network, as shown in Fig. 2(b). For ideal linear modelling elements, the corresponding transformation between hydraulic and mechanical elements are

$$c_1 = {A_P}^2 R_1 \tag{3}$$

$$c_2 = {A_P}^2 R_2 \tag{4}$$

$$b = {A_P}^2 I_2. \tag{5}$$

A schematic plot of a mechanical-hydraulic device in Fig. 3, is proposed to realise the ideal hydraulic network topology in Fig. 2(b). As shown in Fig. 3, the two cylinder chambers are connected with a helical tube, which provides the hydraulic inertance and resistance, and there are valves in the piston providing the resistance. When the piston is moving relatively to the cylinder, two main flow paths, where one is through the helical tube and the other through the valves on the piston, are considered and hydraulically parallel. Note that the physical component nonlinearity and parasitic effects will be introduced in Section 2.2.

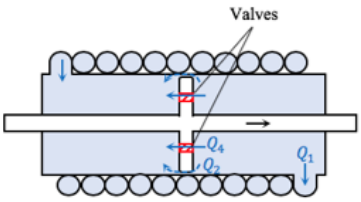
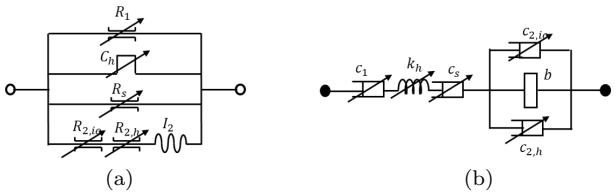


Fig. 3. A schematic plot of a mechanical-hydraulic device which can realise the network in Fig. 2 (Liu et al., 2019).

2.2 Experimental Identification of Component Nonlinearities and Parasitic Effects

As introduced in the previous section, based on the linear assumptions, the mechanical-hydraulic device can realise the linear mechanical network topology shown in Fig. 2(a). In order to achieve the desirable network element properties or making use of them, component nonlinearities and parasitic effects need to be modelled. Note that due to the introduction of nonlinearities, the linear transformation between hydraulic and mechanical elements, such as equations (3), (4) and (5), will not hold. However, the mechanical coupling equations between force and pressure difference, relative velocity and flow rate (equations (1) and (2)), still hold, hence the topological connection of the nonlinear equivalent mechanical network is the same as that of the linear equivalent one.



(a) (b)

Fig. 4. (a) The nonlinear hydraulic network of the mechanical-hydraulic device, (b) the equivalent nonlinear mechanical network of the mechanical-hydraulic device (Liu et al., 2019)

.

Fig. 4(a) shows the nonlinear hydraulic network of the mechanical-hydraulic device, where R_1 is nonlinear orifice

damping, C_h is the hydraulic compliance due to fluid compressibility, R_s is the nonlinear leakage damping, $R_{2,io}$ is the inlet/outlet damping of the helical tube, $R_{2,h}$ is the damping due to surface friction in the tube and I_2 is the tube inertance. Using the mechanical-hydraulic coupling equations (equations (1) and (2)), the nonlinear equivalent mechanical network is presented in Fig. 4(b). Those nonlinear element properties can be identified in separate subsets using designated tests (Liu et al., 2019), which is named as a generalisable model developing approach in Liu et al. (2019).

2.3 Final Nonlinear Mechanical Network-based Model

By implementing all the component nonlinearities and parasitic effects, the final nonlinear mechanical network-based model has been developed, as shown in Fig. 5. In this network model, it has been found that the effects of piston leakage c_s and tube inlet/outlet damping $c_{2,io}$ are negligible. Extra effects, such as the friction force f, coupler stiffness k_s and backlash p are experimentally identified to obtain accurate terminal behaviour. The key element properties (c_1, $c_{2,h}$ and b) in this nonlinear model are expressed by formulae with key design parameters, such as the valve size, tube length and cross-section area (see Table 6 of Liu et al. (2019)). The validity of this nonlinear model has been verified using different design parameter settings.

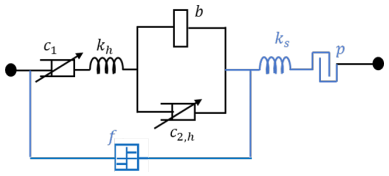


Fig. 5. A schematic plot of a mechanical-hydraulic device which can realise the network in Fig. 2 (Liu et al., 2019).

3. CONCLUSION

Using the feasibility study, a nonlinear mechanical network-based model has been developed for a force-controlling device which could realise the linear mechanical network using nonlinear multidomain components. This modelling approach is applicable for different networks and realisation across different domains.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC) (Grant Reference: EP/P013546/1).

REFERENCES

Chen, M.Z.Q. and Smith, M.C. (2009). Restricted complexity network realizations for passive mechanical control. *IEEE Transactions on Automatic Control*, 54(10), 2290–2301.

Hughes, T.H. (2020). Minimal series-parallel network realizations of bicubic impedances. *IEEE Transactions on Automatic Control*, 1–1. doi: 10.1109/TAC.2020.2968859.

Jiang, J.Z., Matamoros-Sanchez, A.Z., Goodall, R.M., and Smith, M.C. (2012). Passive suspensions incorporating inerters for railway vehicles. *Vehicle System Dynamics*, 50(sup1), 263–276.

Lewis, T., Li, Y., Tucker, G., Jiang, J., Zhao, Y., Neild, S., Smith, M., Goodall, R., and Dinmore, N. (2019). Improving the track friendliness of a four-axle railway vehicle using an inertance-integrated lateral primary suspension. *Vehicle System Dynamics*, 1–20.

Li, Y., Jiang, J.Z., and Neild, S.A. (2019). Optimal fluid passageway design methodology for hydraulic engine mounts considering both low and high frequency performances. *Journal of Vibration and Control*, 25(21-22), 2749–2757.

Liu, X., Jiang, J.Z., Titurus, B., and Harrison, A. (2018). Model identification methodology for fluid-based inerters. *Mechanical Systems and Signal Processing*, 106, 479–494.

Liu, X., Titurus, B., and Jiang, J.Z. (2019). Generalisable model development for fluid-inerter integrated damping devices. *Mechanism and Machine Theory*, 137, 1–22.

Schönfeld, J. (1954). Analogy of hydraulic, mechanical, acoustic and electric systems. *Applied Scientific Research, Section A*, 3(1), 417–450.

Smith, M.C. (2002). Synthesis of mechanical networks: the inerter. *IEEE Transactions on automatic control*, 47(10), 1648–1662.

Smith, M.C. and Wang, F.C. (2004). Performance benefits in passive vehicle suspensions employing inerters. *Vehicle system dynamics*, 42(4), 235–257.

Wang, F.C. and Chan, H.A. (2011). Vehicle suspensions with a mechatronic network strut. *Vehicle System Dynamics*, 49(5), 811–830.

Yamamoto, K. and Smith, M.C. (2015). Bounded disturbance amplification for mass chains with passive interconnection. *IEEE Transactions on Automatic Control*, 61(6), 1565–1574.

Zhang, S.Y., Jiang, J.Z., and Neild, S. (2017a). Optimal configurations for a linear vibration suppression device in a multi-storey building. *Structural Control and Health Monitoring*, 24(3), e1887.

Zhang, S.Y., Jiang, J.Z., and Neild, S.A. (2017b). Passive vibration control: a structure–immittance approach. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2201), 20170011.

Zhang, S.Y., Li, Y.Y., Jiang, J.Z., Neild, S.A., and Macdonald, J.H. (2019a). A methodology for identifying optimum vibration absorbers with a reaction mass. *Proceedings of the Royal Society A*, 475(2228), 20190232.

Zhang, S., Zhu, M., Li, Y., Jiang, J., Ficca, R., Czechowicz, M., Neilson, R., Neild, S., and Herrmann, G. (2019b). Ride comfort enhancement for passenger vehicles using the structure-immittance approach. *Vehicle System Dynamics*, 1–22.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Hyperbolic Secant Varieties of M-Curves

Rainer Sinn * **Mario Kummer** **

* *Universität Leipzig, Leipzig, D-04109 Germany (e-mail: rainer.sinn@uni-leipzig.de).*
** *Technische Universität Dresden, Dresden, D-01062 Germany (e-mail: mario.kummer@tu-dresden.de).*

Abstract: We present recent results joint with Mario Kummer (TU Dresden) on convex hulls of curves. We see a large family of examples where these convex hulls turn out to be hyperbolicity cones. For convex hulls of elliptic curves, we are able to show that these hyperbolicity cones are spectrahedra, generalizing previous results by Henrion and Scheiderer.

Keywords: spectrahedra, convex hull, algebraic curves, hyperbolic polynomial

1. INTRODUCTION

Hyperbolic polynomials and the associated algebraic hypersurfaces are at the intersection of algebraic geometry, optimization, combinatorics, and computer science. Historically, they go back to partial differential equations in the work of Peter Lax and have started a new life in convex programming since the advent of interior point methods. In real algebraic geometry, (smooth) hyperbolic hypersurfaces are an extremal topological type, namely the most nested.

In this paper, we provide a geometric construction of highly singular hyperbolic hypersurfaces, which is interesting from several perspectives. The hyperbolicity of secant varieties of M-curves is closely linked to a property of linear systems on real algebraic curves that we call *vastly real*. This is a common generalization of Ahlfors's circle maps, i.e. real fibered (also called separating or totally real) morphisms to the projective line, and Mikhalkin and Orevkov's maximally writhed links. Vastly real linear systems on curves are an extremal real embedding into projective spaces of almost any odd dimension.

From the point of view of optimization, highly singular hyperbolicity cones are necessary to even have a chance to construct expressive hierarchies analogous to sum-of-squares and moment methods widely used in semidefinite programming (as shown in the work of Saunderson Saunderson (2020) building on Averkov (2019)). Our geometric construction provides such examples. It can be used to write certain convex hulls of connected components of real curves as projections of hyperbolicity cones. The hyperbolic secant hypersurfaces for curves of genus at least 2 are a promising testing ground for the Generalized Lax-Conjecture, which claims that they admit definite determinantal representations up to a cofactor with controlled hyperbolicity cone. Vinnikov (2012). We construct examples of hyperbolic polynomials with a highly singular hyperbolicity cone so that these cones have an interesting facial structure. Our construction also generalizes the Hankel spectrahedron for binary forms to the convex hull of (a connected component) of an M-curve of genus greater than 1. We also show that the hyperbolicity cones constructed in this way are simplicial and therefore do not admit small SDP-lifts. In general, we do not know, if these hyperbolicity cones are spectrahedral.

In classical algebraic geometry, an attractive class of examples of hyperbolic hypersurfaces are definite symmetroids: hypersurfaces whose defining polynomials are the determinant of real symmetric matrix pencils that contain a definite matrix. We show the existence of such determinantal representations for secant varieties of elliptic normal M-curves.

2. CONCLUSION

This work provides many examples of highly singular hyperbolic polynomials. Singularities are necessary for an interesting facial structure and therefore for expressive powers of possible hyperbolic hierarchies. So these examples might lead to interesting applications later. Moreover, they are a testing ground for the generalized Lax conjecture. In genus 1, however, the hyperbolicity cones in question are spectrahedral in the easiest possible way, generalizing previous results by Henrion Henrion (2011) and Scheiderer Scheiderer (2011).

REFERENCES

Averkov, G. (2019). Optimal size of linear matrix inequalities in semidefinite approaches to polynomial optimization. *SIAM J. Appl. Algebra Geom.*, 3(1), 128–151.

Henrion, D. (2011). Semidefinite representation of convex hulls of rational varieties. *Acta Appl. Math.*, 115(3), 319–327.

Saunderson, J. (2020). Limitations on the expressive power of convex cones without long chains of faces. *SIAM J. Optim.*, 30(1), 1033–1047.

Scheiderer, C. (2011). Convex hulls of curves of genus one. *Adv. Math.*, 228(5), 2606–2622.

Vinnikov, V. (2012). LMI representations of convex semialgebraic sets and determinantal representations of algebraic hypersurfaces: past, present, and future. In *Mathematical methods in systems, optimization, and control. Festschrift in honor of J. William Helton*, 325–349. Basel: Birkhäuser.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On computing the H_2 norm using the polynomial Diophantine equation

Timothy H. Hughes and Gareth H. Willetts *

* *University of Exeter, Penryn Campus, Cornwall, U.K. (e-mail: T.H.Hughes@exeter.ac.uk, G.H.Willetts@exeter.ac.uk).*

Abstract: An explicit algorithm will be presented for computing the H_2 norm of a single input single output system from the coefficients in its transfer function. The algorithm follows directly from Cauchy's residue theorem, and the most computationally intensive step involves solving a polynomial Diophantine equation. This can be efficiently solved using subresultant sequences in a fraction-free variant of the extended Euclidean algorithm. The coefficients in these subresultant sequences correspond to the Hurwitz determinants, whereby a stability test can be obtained alongside computing the H_2 norm with little additional computational effort. Implementations of the algorithm symbolically, in exact arithmetic, and in floating-point arithmetic will be presented. The accompanying talk will demonstrate an example application on the design of passive train suspension systems that optimise passenger comfort. The example will demonstrate the algorithm's greater robustness and computational efficiency relative to H_2 norm algorithms requiring the computation of the controllability or observability Gramians. The more general application of the techniques to the realisation of optimal lumped-parameter networks will also be discussed.

Keywords: H_2 optimal control; Mechanical systems; Symbolic computation; Real algebraic geometry; Linear systems; Subresultant sequences.

1. INTRODUCTION

The H_2 norm is a widely used metric for characterising system performance, corresponding to the square root of the power spectral density of the system's output in response to zero mean white noise input of unit power spectral density. It is a natural measure of system performance for the design of mechanical networks, such as vehicle suspension systems; for example, it is commonly used to characterise passenger comfort as a vehicle traverses a rough surface (see, e.g., Wang et al., 2009). The design of the famous Linear Quadratic Gaussian controller also corresponds to a H_2 norm minimisation problem. There is therefore a need for efficient algorithms for the computation of the H_2 norm of a given system. In safety critical applications, or when numerical robustness is a consideration, it can be desirable to compute the H_2 norm using exact arithmetic. Moreover, in the design of lumped-parameter systems, it can be desirable to obtain symbolic expressions for the H_2 norm in terms of the system's parameters. For example, this is useful in the design of optimal mechanical networks, or in other structured H_2 norm optimisation problems, where it is necessary to choose one or more system parameters to optimise a H_2 norm performance measure.

In this paper, an algorithm will be presented for the computation of the H_2 norm of a single input single output system from the coefficients in its transfer function. The most computationally demanding step in the algorithm corresponds to solving a polynomial Diophantine equation. This equation arises from the application of Cauchy's residue theorem to evaluate the frequency domain integral formula (obtained from Parseval's theorem) for the H_2 norm. It will be shown how this equation can be solved efficiently by a fraction-free variant of the extended Euclidean algorithm, which corresponds to the computation of subresultant and remainder polynomials generated from the even and odd part of the denominator polynomial in the system's transfer function. Moreover, the coefficients in these subresultant polynomials correspond to Hurwitz determinants, whereby the stability of the system can be determined alongside the H_2 norm computation. Examples will be presented to demonstrate the robustness and computational efficiency of the algorithm as compared to H_2 norm algorithms that involve computation of the controllability or observability Gramians. The algorithm is broadly applicable to the design of optimal lumped-parameter networks (such as the aforementioned vehicle suspension systems), and more general structured H_2 norm optimisation problems.

The notation employed is as follows. $\mathbb{R}$ denotes the real numbers; $\mathbb{R}^{m\times n}$ the real matrices with m rows and n columns; and $\mathbb{R}[s]$ and $\mathbb{R}(s)$ the univariate polynomials and rational variables in the indeterminate s, respectively. If $p \in \mathbb{R}[s]$, then $\deg(p(s))$ denotes its degree, and $\mathrm{LC}(p(s))$ its leading coefficient. If $G \in \mathbb{R}(s)$, then $\|G\|_2$ denotes its H_2 norm. For a complex number z, its conjugate is denoted z^*, and j denotes the imaginary unit $\sqrt{-1}$. Finally, if $x \in \mathbb{R}$, then $\lceil x \rceil$ rounds $x \in \mathbb{R}$ up to the next integer, and $\lfloor x \rfloor$ rounds $x \in \mathbb{R}$ down to the previous integer; and if x is an integer then $x\%2$ denotes the remainder of x upon division by 2.

2. COMPUTING THE H_2 NORM USING THE POLYNOMIAL DIOPHANTINE EQUATION

The H_2 norm of a linear system has a number of well known equivalent characterisations. From the perspective of system performance, it is most naturally characterised as the power spectral density of the system's output when the input is zero-mean white noise whose power spectral density is equal to the identity matrix. For the purposes of computation, there are three alternative characterisations, corresponding to the system's impulse response, frequency response, and the controllability or observability Gramian. For the case of a (rational and strictly proper) single input single output system with impulse response $g(t)$, frequency response $G(s) \in \mathbb{R}(s)$ (the Laplace transform of $g(t)$), and state-space realization (A, B, C) (i.e., $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times 1}$ and $C \in \mathbb{R}^{1\times n}$ satisfy $G(s) = C(sI - A)^{-1}B$), these three alternative characterisations are as described next. It should be noted that the H_2 norm as defined below requires the system to be asymptotically stable, and accordingly this will be assumed to be the case throughout. The algorithm proposed in this paper for its computation contains a test for stability and can flag when this condition is not met.

Firstly, the H_2 norm is the 2-norm of the impulse response, i.e., $\|G\|_2^2 = \int_{-\infty}^{\infty} g(t)^2 \mathrm{d}t$.

Secondly, using Parseval's theorem, this can be evaluated using the frequency response $G(j\omega)$ as follows:

$$\|G\|_2^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(j\omega)^* G(j\omega)\mathrm{d}\omega. \tag{1}$$

Thirdly, $\|G\|_2^2 = B^T L_o B = C L_c C^T$, where L_o and L_c are the observability and controllability Gramians, which are the solutions to the Lyapunov equations $A^T L_o + L_o A + C^T C = 0$ and $A L_c + L_c A^T + B B^T = 0$, respectively.

Similar characterisations to the above also hold for multi-input multi-output systems (see, e.g., Zhou et al., 1996, pp. 112–113).

Owing to the abundance of efficient computational methods for solving Lyapunov equations, such as those characterising the observability and controllability Gramians, then algorithms for computing the H_2 norm typically employ the third of the aforementioned characterisations. While the Lyapunov equations themselves are linear in the entries in the observability and controllability Gramians, the size of such equations are considerably greater than the state dimension, and the solutions depend in a complicated manner on the entries in the matrices A, B and C. In contrast, in this paper, an algorithm will be presented based on the second of the aforementioned characterisations for the H_2 norm. The most computationally demanding step in the calculation corresponds to solving a structured linear equation of dimension equal to the state dimension. Moreover, the structural properties of this equation can be handled in a computationally efficient manner via a fraction-free variant of the extended Euclidean algorithm.

The algorithm for the computation of the H_2 norm will be stated in terms of the coefficients in the numerator and denominator polynomials of the transfer function:

$$G(s) = \frac{c(s)}{a(s)} = \frac{c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \ldots + c_1 s + c_0}{a_n s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0}, \tag{2}$$

where, without loss of generality, we let $a_n > 0$, and we assume that $c(s)$ and $a(s)$ have no common roots in the closed right half plane (whereupon G is asymptotically stable if and only if the roots of $a(s)$ are all in the open left half plane). Following Ablowitz (2003, pp. 220–221), it follows that the integral in equation (1) is equal to $\oint_D G(-s)G(s)\mathrm{d}s/(2\pi j)$, where D is a contour that traverses the imaginary axis from a point $s = -jR$ to the point $s = jR$, then follows a semicircular arc of radius R into the left half plane, for any given $R > 0$ such that this contour encloses all of the poles of $G(s)$ (see also Zhou et al., 1996).

Since the roots of $a(s)$ are all in the open left half plane, then $a(s)$ and $a(-s)$ have no roots in common, whereupon there exists a unique solution to the polynomial Diophantine equation $c(-s)c(s) = a(s)x(s)+a(-s)y(s)$ for which the degrees of $x(s)$ and $y(s)$ are strictly less than n. Moreover, it can be noted that if the pair $(x(s), y(s))$ satisfies the aforementioned equation, then so too does the pair $(y(-s), x(-s))$, and it follows that $x(s) = y(-s)$. In other words, $y(s)$ is the unique polynomial whose degree is strictly less than n that solves the equation

$$c(-s)c(s) = a(s)y(-s) + a(-s)y(s). \tag{3}$$

Rearranging (3) and substituting it into the previous contour integral yields

$$2\pi j\|G\|_2^2 = \oint_D \frac{c(s)c(-s)}{a(s)a(-s)}\mathrm{d}s = \oint_D \frac{y(-s)}{a(-s)}\mathrm{d}s + \oint_D \frac{y(s)}{a(s)}\mathrm{d}s.$$

Since the roots of $a(s)$ are all in the open left half plane, then D contains all of the poles of $y(s)/a(s)$ and none of the poles of $y(-s)/a(-s)$, whereupon by Cauchy's residue theorem it follows that

$$\|G\|_2^2 = \frac{1}{2\pi j}\oint_D \frac{y(s)}{a(s)}\mathrm{d}s.$$

Since D contains all of the poles of $y(s)/a(s)$, then the above contour integral can be evaluated using the concept of the *residue at infinity* (see Ablowitz, 2003, pp. 211–212). Specifically, from (Ablowitz, 2003, equations (4.1.13) and (4.1.14)), it follows that

$$\|G\|_2^2 = \lim_{s\to\infty}\left(\frac{sy(s)}{a(s)}\right) = \frac{y_{n-1}}{a_n},$$

where y_{n-1} denotes the coefficient of s^{n-1} in $y(s)$. In summary, the computation of the H_2 norm of the system whose transfer function is as in (2) amounts to the determination of the coefficient y_{n-1} from the polynomial solution $y(s)$ (of degree $n-1$ or less) to equation (3).

Now, let

$$\begin{aligned} c^e(s) &= c_0 + c_2 s + \ldots + c_{2\lfloor\frac{n-1}{2}\rfloor} s^{\lfloor\frac{n-1}{2}\rfloor}, \\ c^o(s) &= c_1 + c_3 s + \ldots + c_{2\lceil\frac{n-3}{2}\rceil+1} s^{\lceil\frac{n-3}{2}\rceil}, \\ \text{and } z_0(s) &= (c^e(s))^2 - s(c^o(s))^2, \end{aligned} \tag{4}$$

whereupon $c(s) = c^e(s^2) + sc^o(s^2)$ and

$$c(-s)c(s) = (c^e(s^2) - sc^o(s^2))(c^e(s^2) + sc^o(s^2)) = z(s^2).$$

Further, let

$$\begin{aligned} a^e(s) &= a_0 + a_2 s + \ldots + a_{2\lfloor\frac{n}{2}\rfloor} s^{\lfloor\frac{n}{2}\rfloor}, \\ \text{and } a^o(s) &= a_1 s + a_3 s^2 + \ldots + a_{2\lceil\frac{n}{2}\rceil-1} s^{\lceil\frac{n}{2}\rceil}, \end{aligned} \tag{5}$$

whereupon $a(s) = a^e(s^2) + \frac{1}{s}a^o(s^2)$. It can be shown that $a^e(s)$ and $a^o(s)$ do not share any common roots since the

roots of $a(s)$ are all in the open left half plane, and it follows that there exist unique polynomials $f(s)$ and $g(s)$ such that the degree of $f(s)$ (resp. $g(s)$) is strictly less than the degree of $a^o(s)$ (resp. $a_e(s)$) and

$$a^e(s)f(s) + a^o(s)g(s) = z_0(s). \tag{6}$$

Then, with the notation

$$y(s) = \tfrac{1}{2}(f(s^2) - sg(s^2)), \tag{7}$$

it is easily shown that the degree of y is strictly less than n, and that $a(s)y(-s) + a(-s)y(s) = c(-s)c(s)$. In other words, $y(s)$ in equation (7) is the unique solution to equation (3) for which the degree of $y(s)$ is strictly less than n, whereupon the coefficient y_{n-1} of s^{n-1} in the polynomial $y(s)$ is determined from the solutions $f(s)$ and $g(s)$ to the polynomial Diophantine equation (6).

3. EFFICIENT COMPUTATION OF THE SOLUTION TO THE POLYNOMIAL DIOPHANTINE EQUATION

It has been shown that the H_2 norm of the system whose transfer function is as in (2) is equal to y_{n-1}/a_n, where y_{n-1} is the coefficient of s^{n-1} in the polynomial $y(s)$ in (7), where $f(s)$ and $g(s)$ are the solutions to the polynomial Diophantine equation (6). Here, $a_e(s)$, $a_o(s)$ and $z_0(s)$ are directly determined from the coefficients $c_0, c_1, \ldots c_{n-1}$ and $a_0, a_1, \ldots, a_n$ in the transfer function $G(s)$ using equations (4) and (5). In this and the next section, efficient algorithms for the computation of the solution to this Diophantine equation will be presented.

First, by equating coefficients of s in equation (6), it follows that the coefficients in the polynomials $f(s)$ and $g(s)$ can be obtained by finding the solution x to the linear equation

$$H^T x = b, \tag{8}$$

where H is the $n \times n$ Hurwitz matrix

$$H = \begin{bmatrix} a_{n-1} & a_{n-3} & a_{n-5} & \cdots \\ a_n & a_{n-2} & a_{n-4} & \cdots \\ 0 & a_{n-1} & a_{n-3} & \cdots \\ 0 & a_n & a_{n-2} & \cdots \\ & & & \ddots \end{bmatrix}, \tag{9}$$

and $b^T = [z_{0,n-1}\ z_{0,n-2}\ \cdots\ z_{0,1}\ z_{0,0}]$ where $z_0(s) = z_{0,n-1}s^{n-1} + z_{0,n-2}s^{n-2} + \ldots + z_{0,1}s + z_{0,0}$. Here, if n is odd (resp., even), then the odd (resp., even) entries in the solution x to equation (8) correspond to the coefficients in $f(s)$ in equation (6), and the even (resp., odd) entries correspond to the coefficients in $g(s)$, in descending degree. In either case, it follows that $y_{n-1} = (-1)^{n+1}x_1/2$, where x_1 denotes the first entry in the solution x to equation (8).

The preceding characterisation of the H_2 norm is similar to the approach taken by Betser et al. (1995) to characterise the solution P to the Lyapunov equation $-PA - A^TP = Q$ in the case that A is a companion matrix. In contrast to the preceding derivation, the result of Betser et al. (1995) used the theory of matrix polynomials. A similar approach was followed by Hughes (2016), where an alternative characterisation was also obtained in terms of the Bezoutian of the polynomials $a^o(s)$ and $a^e(s)$, which allows one to exploit the algorithms of Bini and Gemignani (1998) for efficient triangularisation of Bezoutians.

This solution method via the polynomial Diophantine equation (6) also lends itself to both exact computation over the integers and to symbolic computation. This is of particular interest in safety critical applications, and in structured H_2 norm optimisation problems. In these cases, the solution can be efficiently obtained via the algorithm to be described in the next section.

4. EXACT OR SYMBOLIC COMPUTATION USING SUBRESULTANT SEQUENCES

In this section, an algorithm for solving the polynomial Diophantine equation (6) will be provided that is particularly suitable for exact or symbolic computation. The algorithm amounts to the computation of subresultant and remainder sequences in a generalised version of the extended Euclidean algorithm (see, e.g., Basu et al., 2006, Section 8.3). The coefficients of these subresultant and remainder sequences correspond to subdeterminants of the Hurwitz matrix, and so a stability test can be performed with little added computational cost. The method is also fraction-free (i.e. at each stage of the computation, the results are integers whenever the coefficients in the originating transfer function are integers), which allows for efficient exact or symbolic computation. The algorithm and its properties are presented in this section. The proof will follow in a journal paper currently in preparation.

First, note that the definitions for the polynomials change depending on whether the degree of the transfer function, n, is odd or even. Owing to space constraints, we consider only the case where n is odd in the following. Define

$$p_1(s) = a^o(s) = a_1 s + a_3 s^2 + \ldots + a_{2\lceil \frac{n}{2} \rceil - 1} s^{\lceil \frac{n}{2} \rceil},$$
$$p_2(s) = a^e(s) = a_0 + a_2 s + \ldots + a_{2\lfloor \frac{n}{2} \rfloor} s^{\lfloor \frac{n}{2} \rfloor},$$

and recall the definition of $z_0(s)$ in terms of the coefficients in the numerator polynomial of the transfer function $G(s)$ from equation (4). To calculate the H_2 norm, we recursively compute polynomials $z_1, \ldots, z_{n-1}$ and $p_3, \ldots, p_{n+1}$ as follows. Firstly, the polynomials $z_1(s)$ and $p_3(s)$ are obtained thus:

$$z_1(s) = \mathbb{LC}(p_1)z_0(s) - s^{\lceil \frac{n-2}{2} \rceil} z_{0,n-1}p_1(s),$$
$$p_3(s) = \mathbb{LC}(p_2)p_1(s) - s^{n\%2}\mathbb{LC}(p_1)p_2(s).$$

Here, and hereafter, $z_{i-1,n-i}$ denotes the coefficient of degree $n-i$ in the polynomial $z_{i-1}(s)$, which could equal 0 ($i = 1, 2, \ldots, n$). Next, the polynomials $z_2(s)$ and $p_4(s)$ are obtained thus:

$$z_2(s) = \mathbb{LC}(p_2)z_1(s) - s^{\lceil \frac{n-3}{2} \rceil} z_{1,n-2}p_2(s),$$
$$p_4(s) = \mathbb{LC}(p_3)p_2(s) - s^{(n-1)\%2}\mathbb{LC}(p_2)p_3(s).$$

Then, given polynomials $p_i(s), p_{i+1}(s)$ and $z_{i-1}(s)$, the polynomials $p_{i+2}(s)$ and $z_i(s)$ are obtained thus:

$$z_i(s) = \frac{\mathbb{LC}(p_i)}{\mathbb{LC}(p_{i-1})} z_{i-1}(s) - s^{\lceil \frac{n-i-1}{2} \rceil} \frac{z_{i-1,n-i}}{\mathbb{LC}(p_{i-1})} p_i(s),$$
$$p_{i+2}(s) = \frac{\mathbb{LC}(p_{i+1})}{\mathbb{LC}(p_{i-1})} p_i(s) - s^{(n-i+1)\%2} \frac{\mathbb{LC}(p_i)}{\mathbb{LC}(p_{i-1})} p_{i+1}(s),$$

for $i = 3, 4, \ldots, n-1$. The coefficients $\mathbb{LC}(p_k)$ correspond to principal subdeterminants of the Hurwitz matrix, whereby $\mathbb{LC}(p_k) > 0$ for all $k = 0, \ldots, n+1$ due to $G(s) = c(s)/a(s)$ being asymptotically stable (Gantmacher, 1980, p. 194). Moreover, if any one of these coefficients is non-positive, then the algorithm can terminate with notifica-

tion that the roots of $a(s)$ are not all in the open left-half plane and thus the H_2 norm is ill-posed. Asymptotic stability is also sufficient to guarantee that the degrees of the polynomials satisfy $\deg(p_{i+2}(s)) = \deg(p_i(s)) - 1$ for $i = 1, 2, \ldots, n-1$, and it is clear that $\deg(z_i(s)) < \deg(z_{i-1}(s))$ for $i = 1, 2, \ldots, n-1$, whereupon it follows that $z_{n-1}(s)$ and $p_{n+1}(s)$ are scalar constants.

Having computed this sequence of polynomials, the H_2 norm is then obtained as

$$\frac{z_{n-1}}{2a_n p_{n+1}}.$$

It can be shown that the coefficients of the polynomials $p_{i+2}(s)$ and $z_i(s)$ will be integers whenever the coefficients of the originating polynomials p_1, p_2 and z_0 are integers, for $i = 1, 2, \ldots, n-1$. In this case, all quantities can be calculated exactly and efficiently using integer arithmetic. This property also facilitates efficient symbolic computation when the coefficients are parametric, for example when considering lumped parameter networks such as mechanical suspension systems.

The performance of this algorithm has been compared to the `norm` command in MATLAB 2021b and is of the order of 100 times faster for transfer functions of degrees in the range from 5 to 21. MATLAB tends to take longer to execute code when running code for the first time, while it loads required functions and allocates memory for variables. This is illustrated in Figure 1; it is particularly prominent over the first 10 runs, but there are still spikes in execution time during the first 1000. These spikes in execution time still exist in runs after the first 1000 but are less prominent.

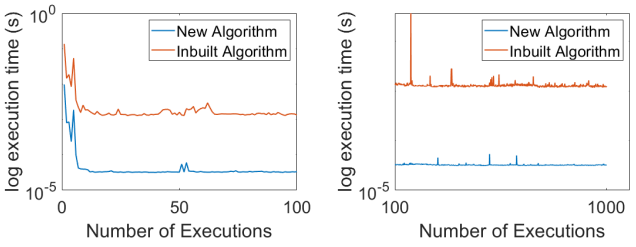


Fig. 1. Execution times plotted for the first 100 and following 1000 executions, respectively, for a transfer function with degree $n = 21$.

The following table presents the executation times for this algorithm in comparison with the `norm` command. Owing to the speed up in performance, these results represent the average execution time between the 1000th and 10000th run. A similar order of magnitude improvement in execution time is also observed for the first run and the first 1000 runs. These metrics were obtained on an Intel i7-12700K, and are reported in microseconds to 3 significant figures. Transfer functions were constructed as follows: $c(s) = \sum_{i=0}^{n-1} s^i$, $a(s) = \sum_{i=0}^{n} \binom{n}{i} s^i$.

n	MATLAB (inbuilt)	MATLAB (New Algorithm)
5	707	9.09
7	742	14.6
9	784	14.9
21	1200	32.3

Furthermore, the command `isstable`, used to test the stability of a transfer function in MATLAB 2021b, adds a further 205 microseconds to the computation time for $n = 21$, in comparison to checking the leading coefficient of $p_i > 0$, which has negligible effect on computation time.

In addition, direct solution to equation (8) was tested, using the backslash operator in MATLAB. This was roughly 3 times slower than the method outlined in this paper, yet still significantly faster than MATLAB's inbuilt function, averaging at 98.3 microseconds between the 1000th and 10000th runs for $n = 21$.

The algorithm has particular value in terms of optimisation of lumped-parameter networks, such as mechanical networks. For example, many vehicle suspension design problems can be characterised in terms of optimising the network parameters (such as the spring stiffnesses, damping rates and inertias) in order to optimise the H_2 norm of some driving-point or transfer admittance (see, e.g., Wang et al., 2009). The driving-point or transfer admittance will be functions of these design parameters whose coefficients can be obtained directly using graph theoretic methodologies (see, e.g., Percival, 1953). Optimisation of the H_2 norm can be implemented efficiently with the new algorithm, either numerically or symbolically. The symbolic approach also enables computation of the gradient vector and Hessian matrix to allow application of efficient interior-point or similar methodologies for optimisation.

REFERENCES

Ablowitz, M. (2003). *Complex Variables: Introduction and Applications*. Cambridge University Press.

Basu, S., Pollack, R., and Roy, M. (2006). *Algorithms in Real Algebraic Geometry*. Springer.

Betser, A., Cohen, N., and Zeheb, E. (1995). On solving the Lyapunov and Stein equations for a companion matrix. *Systems and Control Letters*, 25, 211–218.

Bini, D. and Gemignani, L. (1998). Fast fraction-free triangularization of bezoutians with applications to subresultant chain computation. *Linear Algebra Appl.*, 284, 19–39.

Gantmacher, F.R. (1980). *The Theory of Matrices*, volume II. New York : Chelsea.

Hughes, T. (2016). Behavioral realizations using companion matrices and the Smith form. *SIAM Journal on Control and Optimization*, 54(2), 845–865.

Percival, W. (1953). Solution of passive electrical networks by means of mathematical trees. *Proceedings of the IEE - Part III: Radio and Communication Engineering*, 100(65), 143–150.

Wang, F., Liao, M., Liao, B., Su, W., and Chan, H. (2009). The performance improvements of train suspension systems with mechanical networks employing inerters. *Vehicle System Dynamics*, 47(7), 805–830.

Zhou, K., Doyle, J., and Glover, K. (1996). *Robust and Optimal Control*. New Jersey : Prentice Hall.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Optimum Mechanical Network Identification Methodologies: With and Without Mass Elements ⋆

Yi-Yuan Li * **Sara Ying Zhang** ** **Jason Zheng Jiang** *
Simon Neild * **John Macdonald** *

* *School of Civil, Aerospace and Mechanical Engineering, University of Bristol, Bristol, BS8 1TR, UK (e-mail: yiyuan.li@bristol.ac.uk, z.jiang@bristol.ac.uk, simon.neild@bristol.ac.uk, john.macdonald@bristol.ac.uk)*
** *Institute of Urban Smart Transportation and Safety Maintenance, Shenzhen University, Shenzhen 518060, People's Republic of China (e-mail: Sara.zhangying@szu.edu.cn)*

Abstract: Traditional linear passive vibration-absorber networks, such as the tuned mass damper (TMD), often contain springs, dampers and masses. Recently there has been a growing trend to supplement or replace the masses with inerters. When considering the absorbers without a mass, a structure-immittance approach was proposed to identify possible configurations consisting of springs, dampers and inerters. This approach can characterise the full class of network layouts with pre-determined numbers of each element type, and also prescribe the allowed value range for each element. More recently, a mass-included passive absorber, the tuned-mass-damper-inerter, was introduced, showing significant performance benefits on vibration suppression. With the aim to further explore the potential of numerous mass-included passive absorber layouts, a more generalised methodology was developed. Using this methodology, a full class of absorber layouts with a mass and a pre-determined number of inerters, dampers and springs connected in series and parallel can be systematically investigated. A 3-storey building model is used to demonstrate the advantages of the proposed approaches for the cases without and with a mass, where the performance improvements can be up to 21.6% and 65.6%, respectively, compared to the TMD.

Keywords: Passive vibration damper, Network topologies, inerter, systematic methodology, reaction mass.

1. INTRODUCTION

Tuned mass dampers (TMDs), in which a reaction mass is attached to a structural system via a spring–parallel–damper connection, are commonly used in a wide range of applications to suppress deleterious vibrations. The inerter was introdued by Smith (2002), which allows electrical networks to be translated over to mechanical ones in a completely analogous way, with the mechanical elements springs, dampers and inerters corresponding to the electrical elements inductors, resistors and capacitors. Hence, any positive-real functions can be synthesised by passive mechanical networks consisting of dampers, inerters and springs using the theorem proposed by Bott and Duffin (1949). When considering of possible configurations with these elements broadly, two approaches are normally used: one structure-based and one immittance-based. Both approaches have their advantages and disadvantages. Later, a new approach was proposed by Zhang et al. (2017) – the structure–immittance approach. Using this approach, a full set of possible series–parallel networks with predetermined numbers of inerters, dampers and springs can be represented by structural immittances, obtained via a proposed general formulation process. Using the structural immittances, both the ability to investigate a class of absorber possibilities together (advantage of the immittance-based approach), and the ability to control the complexity, topology and element values (advantages of the structure-based approach) are provided.

Recently, by adding an inerter into the TMD and employing two structural attachment points, Marian and Giaralis (2014) proposed a new vibration suppression device, the tuned-mass-damper-inerter (TMDI), with significant benefits obtained. It demonstrates the potential advantages of passive vibration absorbers including the mass and the two-terminal mechanical elements springs, dampers and inerters. However, there are countless topological connection possibilities with these elements included, some of which can potentially provide much more advantageous performance. Therefore, an approach was proposed by Zhang et al. (2019) to systematically characterise a full set of passive vibration absorbers with series–parallel connections of a mass and a pre-determined number of springs, dampers and inerters.

⋆ This is a resubmission of the MTNS2020 extended abstract.

In this extended abstract, the structure-immittance and mass-included synthesis approaches are introduced in Sections 2&3, respectively, to demonstrate the construction of the generic networks and their corresponding mathematical representations. Examples containing one spring, one damper and one inerter (termed as 1k1c1b) are demonstrated accordingly. Finally, a three-storey building model is used in Section 4 to show the advantages of the proposed two approaches on identifying beneficial 1k1c1b mechanical networks with and without a mass element.

2. STRUCTURE-IMMITTANCE APPROACH

The structure-immittance approach is based on the force-current analogy where two-terminal networks are considered. Therefore, three elements p, q, r are considered. They can represent in any order either inerters, dampers and springs in the mechanical domain or capacitors, inductors and resistors in the electrical one. Assume the numbers of p, q and r elements are P, Q and R, respectively and $P \leq Q \leq R$ is satisfied by selecting the p, q and r elements appropriately. By using the structure-immittance approach, generic networks which contain explicit information of all topology possibilities for a given number of each component are first constructed. The formulation of a generic network representing P p, Q q and R r elements requires a series of steps, summarised as a flow chart in Figure 1. Detailed steps to obtain the generic networks can be found in the paper by Zhang et al. (2017).

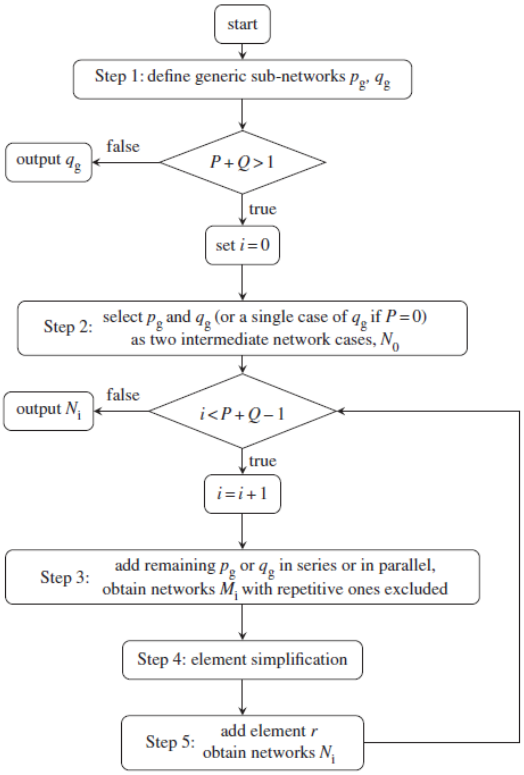


Fig. 1. Flow chart summarising the steps to obtain generic networks done by Zhang et al. (2017) for P p, Q q and R r case where $P \leq Q \leq R$.

One mechanical network example with one spring, one damper and one inerter (i.e. 1k1c1b case) are constructed based on the procedures shown in Figure 1. Two generic networks, termed Q_1 and Q_2, can be obtained as shown in Figure 2. Here we choose the damper and inerter as the base element for the sub-networks and the spring is appointed as the added element. Totally 8 layouts are included by these two generic networks which contain all the possible layouts for the 1k1c1b case.

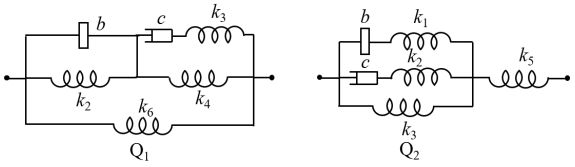


Fig. 2. The generic networks obtained for the 1k1c1b case where only one spring is actually present in each of the generic networks.

After obtaining the generic networks, their corresponding structural-immittances can be derived, which are defined as the transfer functions of generic networks from force to velocity, i.e.

$$Y(s) = \frac{F(s)}{V(s)} \tag{1}$$

$F(s)$ and $V(s)$ are the force and relative velocity across the two terminals in the Laplace domain, respectively. s is the complex frequency parameter of the Laplace transform. Here, for the 1k1c1b case, the structural-immittances of generic networks Q_1 and Q_2 are represented as:

$$\begin{aligned} Y_1(s) &= \frac{bcs^2 + b(k_4+k_6)s + c(k_2+k_6)}{bc(1/k_3)s^3 + bs^2 + cs + k_2 + k_4}, \\ Y_2(s) &= \frac{bc(1/k_1+1/k_2)s^3 + bs^2 + cs + k_3}{b(1/k_1+1/k_5)s^3 + c(1/k_2+1/k_5)s^2 + s} \end{aligned} \tag{2}$$

Note that, for $Y_1(s)$, only one of k_2, $1/k_3$, k_4 and k_6 is positive and all the others are equal to zero. Similarly, for $Y_2(s)$, only one of $1/k_1$, $1/k_2$, k_3 and $1/k_5$ is positive and all the others are equal to zero. These transfer functions can be used, along with the specific constraints on the number of spring elements, to find the optimum mechanical network for a given system and objective functions.

3. MASS-INCLUDED SYNTHESIS APPROACH

The structure-immittance approach, which employed network synthesis theory, made use of the fact that the inerters, dampers and springs all have two terminals. When a reaction mass is included into the networks, a systematic approach becomes much more challenging, as the reaction mass is always regarded as a one-terminal element. In order for network synthesis to be directly applicable to the systematic enumeration of vibration absorbers with a reaction mass, it is necessary to treat the mass as a special two-terminal element, with one terminal notionally connected to the ground, denoted as a notional-ground (NG). Accordingly, terminals connected to physical attachments are denoted as physical-terminals (PTs).

Considering the fact that most vibration suppression devices have no more than two attachment points, investigation of the mass-included synthesis approach focuses on 1PT1NG and 2PT1NG networks. The 1PT1NG network, represented by its force–velocity transfer function $H(s) = F_1(s)/V_1(s)$, is shown in Figure 3(a). At the PT1, the force f_1 ($F_1(s)$ in the Laplace domain) is applied

and results in a velocity v_1 ($V_1(s)$). Figure 3(b) shows a 2PT1NG network, with forces f_1, f_2 and velocities v_1, v_2 at two PTs. Note that because of the presence of a reaction mass, in contrast to the 2PT network (whose immittance function is $Y(s) = F(s)/(V_1(s) - V_2(s))$, see Figure 3(c)), the forces f_1, f_2 of the 2PT1NG network are not equal and opposite. In the following parts, constructions of 1PT1NG and 2PT1NG networks are briefly introduced, and generic networks of an example mechanical network containing a mass and 1k1c1b are demonstrated. Details can be found in the work done by Zhang et al. (2019).

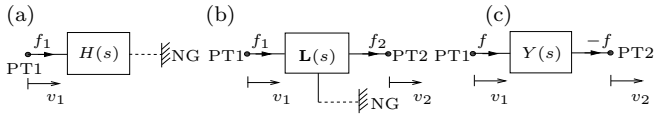


Fig. 3. (a) 1PT1NG network, (b) 2PT1NG network and (c) 2PT network (by Zhang et al. (2019)).

3.1 1PT1NG network layout enumeration

To construct the 1PT1NG networks, we consider joining a 2PT network with a 1PT1NG Immittance-Function-Network (IF-Network), where an IF-Network refers to a network layout with its 2PT sub-networks represented by Immittance-Function-Blocks (IF-Blocks). In order to formulate series-parallel 1PT1NG IF-Networks, a collection of finite numbers of IF-Blocks combined with a mass is considered. A non-unique connection sequence is proposed, with which all possible 1PT1NG IF-Networks can be obtained. Start with a single IF-Block, it can be connected in series or in parallel with other IF-Blocks, however, these always reduce to a single IF-Block. At a certain step, the resulting IF-Block is connected to the mass, where only a series connection is possible, resulting in a new 1PT1NG network. Further addition of IF-Blocks can only be connected in series with this 1PT1NG network, as a parallel connection would necessitate a NG being connected with a PT. Hence, all the IF-Networks can be represented by the generic IF-Network shown in Figure 4 with a single IF-Block $Y(s)$.

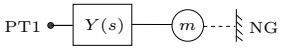


Fig. 4. The generic 1PT1NG IF-Network got by Zhang et al. (2019).

For the 1PT1NG network layouts with one reaction mass and 1k1c1b elements, all network possibilities can be obtained using the generic IF-Network in Figure 4, by enumerating the full class of 2PT network possibilities consisting of 3 elements in $Y(s)$. To this end, the structure-immittance approach can be directly applied to obtain the structural immittance $Y(s)$, which is shown in Eq. 2 for the 1k1c1b case. The transfer function of the generic IF-Network can be obtained as:

$$H(s) = Y(s)ms/(ms + Y(s)) \tag{3}$$

With $H(s)$, the optimum 1PT1NG vibration absorber can be identified for a given vibration suppression problem.

3.2 2PT1NG network layout enumeration

In order to formulate 2PT1NG IF-Networks, a sequence of steps is introduced based on the work by Nishizeki and Saito (1974), as shown in Figure 5. The construction of a three-terminal series-parallel network begins with an empty graph as shown in Step 1. We first consider joining the two terminals, PT1 and NG. Using the generic 1PT1NG IF-Network obtained in the previous section (Figure 4), a network shown in Step 2 is obtained. Then a single IF-Block $Y_2(s)$ is added to the network, resulting in Step 3 via a parallel connection. Consider adding the next IF-Block, $Y_3(s)$, resulting in the new IF-Network shown in Step 4 using the series connection. Note that all the other connection possibilities between Step 3 and the IF-Block $Y_3(s)$ can all be simplified to Step 3. At this point, only a parallel IF-Block can be added, with the resulting IF-Network shown in Step 5, since series IF-Blocks can be reduced to the network of Step 4. Following this parallel addition, only series additions modify the network. Note that series connections at both PT terminals need to be considered, and we define connecting to PT2 as Step 6. An additional IF-Block is then added in series at PT1 – the resultant network is shown in Step 7. Consequent steps will be adding IF-Blocks in parallel then in series by repeating Steps 5-7, until all IF-Blocks in the original collection are used.

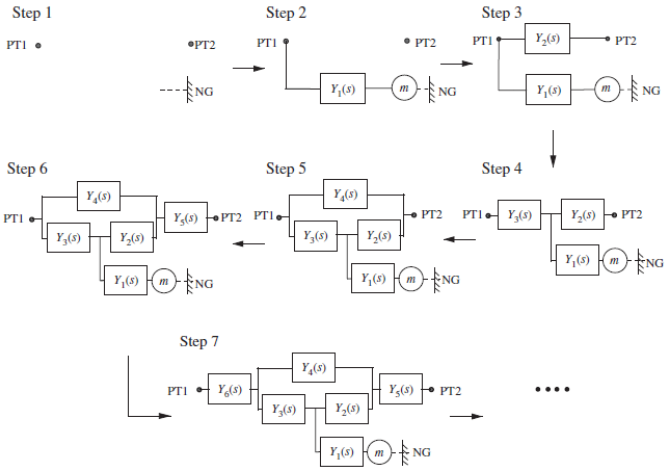


Fig. 5. A procedure summarised by Zhang et al. (2019) to form all possible series-parallel 2PT1NG IF-Networks with a mass and a pre-determined number of IF-Blocks .

Based on the obtained series-parallel 2PT1NG IF-Networks, the rest part focuses on generating generic 2PT1NG IF-Networks. Different from 1PT1NG networks, where one generic IF-Network is sufficient (Figure 4), for 2PT1NG networks, different generic IF-Networks are needed depending on the number, R, of IF-Blocks which are present. In this extended abstract, 1k1c1b case is considered for the possible 2PT1NG generic IF-Networks where at most 3 IF-Blocks exist. Therefore, three generic IF-Networks for R =1, 2, 3 are obtained as shown in Figure 6.

To describe the relations between the velocities and the forces of the 2PT1NG generic IF-Networks in the Laplace domain, an Immittance-Function-Matrix (IF-Matrix), de-

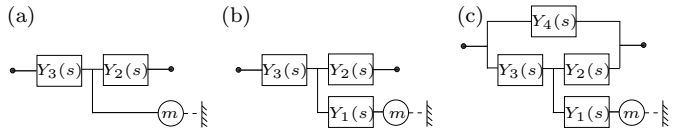


Fig. 6. Generic IF-Networks of 2PT1NG networks containing one spring, one damper and one inerter where the present IF-Blocks (a) $R = 1$, (b) $R = 2$ and (c) $R = 3$.

noted as $\mathbf{L}(s)$, is required, where the velocities and the forces are related in Laplace domain, as:

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \mathbf{L}(s) \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \tag{4}$$

$\mathbf{L}(s)$ is a 2×2 matrix. For generic IF-Networks of 1k1c1b 2PT1NG networks as shown in Figure 6. Their IF-Matrices can be found in the paper by Zhang et al. (2019). With the obtained IF Matrices, the optimum 2PT1NG network containing 1k1c1b with a reaction mass can be identified for a given vibration suppression problem.

4. NUMERICAL APPLICATION ON THE EXAMPLE STRUCTURE

In order to demonstrate the vibration suppression abilities of the mechanical networks identified by the above introduced systematic approaches for cases with and without a mass element, a three degrees-of-freedom (DOFs) structure model is introduced, with floor masses $m_s = 10000$ kg and inter-storey stiffness $k_s = 15000$ kN/m. The structural damping is taken to be zero as it is typically negligible compared with that introduced by absorbers. The absorber is connected between the second and third floors with the reaction mass m =1000 kg if applicable. A brace stiffness $k_b = 0.2k_s$ in series with the absorber and connecting to the lower floor is included. In this example, the inter-storey drift displacement is taken as the performance measure. Therefore the objective function is defined as:

$$J_d = \max_{(\text{over } i)} \left(\max_{(\text{over } \omega)} |T_{s^2 X_0 \to X_{d_i}}| \right), \quad (i = \text{I, II, III}), \tag{5}$$

where $T_{s^2 X_0 \to X_{d_i}}$ denotes the transfer function from the earthquake acceleration to the inter-storey drift displacements and $\text{Max}|T_{s^2 X_0 \to X_{d_i}}|$ is the maximum magnitude of $T_{s^2 X_0 \to X_{d_i}}$.

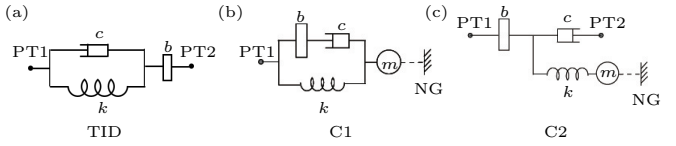


Fig. 7. Optimal mechanical layouts containing one mass and 1k1c1b for the (a) 2PT network (TID), (b) 1PT1NG network (C1) and (c) 2PT1NG network (C2).

When considering 2PT networks, the structure-immittance approach is employed, where the optimum mechanical absorber is the tuned inerter damper (TID) as shown in Figure 7(a), with the parameter values in Table 1. For the 1PT1NG and 2PT1NG networks, generic networks in Figure 4 and Figure 6, respectively, are employed, and the optimum layout is shown in Figure 7 as C1 and C2. Their corresponding parameter values are shown in Table 1. It can be obtained that the performance improvements can be up to 21.6%, 34.8% and 65.6% for the identified mechanical networks of 2PT, 1PT1NG and 2PT1NG cases. Their corresponding frequency domain responses are illustrated in Figure 8.

Table 1. Optimisation results for the optimum 2PT, 1PT1NG and 2PT1NG layouts with $m =$ 1000 kg where applicable.

Network type	TMD	TID	C1	C2
J_d ($\times 10^{-3}$ s^2)	25.0	19.6	16.3	8.60
Percentage Imp. (%)	(–)	(21.6%)	(34.8%)	(65.6%)
Stiffness (kN/m)	2.75×10^2	4.35×10^3	4.92×10^2	2.13×10^2
Damping (kNs/m)	11.7	7.50×10^2	30.5	3.26×10^2
Inertance (kg)	–	7.94×10^3	1.05×10^3	8.10×10^3

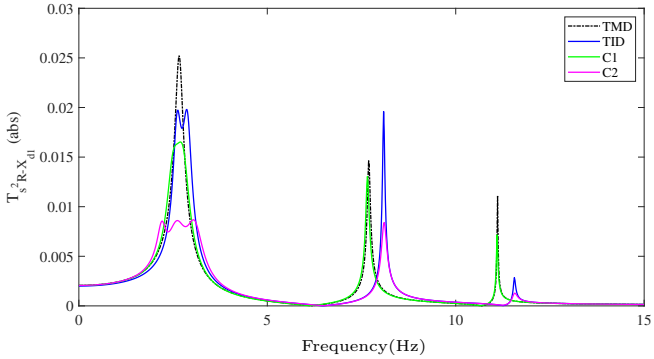


Fig. 8. Frequency response of the inter-storey drift between the ground and the first floor for the TMD (black dashed), TID (blue solid), C1 (green solid) and C2 (magenta solid).

ACKNOWLEDGEMENTS

The authors would like to thank the support of the EPSRC. Jason Zheng Jiang is support by the EPSRC First Grant EP/P013546/1.

REFERENCES

Bott, R. and Duffin, R. (1949). Impedance synthesis without use of transformers. *J. of Appl. Phys.*, 20(8), 816.

Marian, L. and Giaralis, A. (2014). Optimal design of a novel tuned mass-damper-inerter (tmdi) passive vibration control configuration for stochastically support-excited structural systems. *Probabilist. Eng. Mech.*, 38, 156–164.

Nishizeki, T. and Saito, N. (1974). Necessary and sufficient condition for a graph to be three-terminal series-parallel. *IEEE Trans. Circuit Syst.*, 22, 648–653.

Smith, M. (2002). Synthesis of mechanical networks: the inerter. *IEEE Trans. Auto. Contr.*, 47(10), 1648–1662.

Zhang, S.Y., Jiang, J.Z., and Neild, N.A. (2017). Passive vibration control: a structure-immittance approach. *P Roy. Soc. A - Math. Phy.*, 473(2201), 20170011.

Zhang, S.Y., Li, Y.Y., Jiang, J.Z., Neild, N.A., and Macdonald, J. (2019). A methodology for identifying optimum vibration absorbers with a reaction mass. *P Roy. Soc. A - Math. Phy.*, 475(2228), 20190232.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Resource allocation in open multi-agent systems: an online optimization approach

Charles Monnoyer de Galland * **Renato Vizuete** **,*** **Julien M. Hendrickx** * **Paolo Frasca** *** **Elena Panteley** **

* *ICTEAM Institute, UCLouvain, B-1348, Louvain-la-Neuve, Belgium (e-mail: {charles.monnoyer,julien.hendrickx}@uclouvain.be)*
** *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France, (e-mail: {renato.vizuete,elena.panteley}@l2s.centralesupelec.fr)*
*** *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France, (e-mail: paolo.frasca@gipsa-lab.fr)*

Abstract: The resource allocation problem consists in the optimal distribution of a budget between a group of agents. We consider a version of this optimization problem in open systems where agents can be replaced, resulting in variations of the budget and the total cost function to be minimized. We analyze the performance of the Random Coordinate Descent algorithm (RCD) in that setting using natural performance indexes which are related to those used in online optimization. We show that, in a simple setting, both the accumulated error obtained from using the RCD as compared to the optimal solution and the accumulated gain obtained from using the RCD instead of not collaborating grow linearly with the number of iterations considered for the computation, so that in expectation an error cannot be avoided, but remains bounded.

Keywords: Online optimization, Multi-agent systems, Convex optimization, Optimization and control of large-scale network systems, Open systems, Dynamic Resource Allocation.

1. INTRODUCTION

We consider the optimal resource allocation problem, where a fixed amount of resource must be distributed among n agents while minimizing some separable cost function f. Such problems arise in different fields of research, such as actuator networks (Teixeira et al., 2013) or energy resources (Dominguez-Garcia et al., 2012). In some of their formulations, each agent i holds a quantity d_i (which we call here the "demand" of agent i), so that the total amount of resource to be distributed is $\sum_{i=1}^n d_i$; the problem can then be written as

$$\min_{x\in\mathbb{R}^{np}} f(x) = \sum_{i=1}^n f_i(x_i) \quad \text{s.t.} \quad \sum_{i=1}^n x_i = \sum_{i=1}^n d_i, \qquad (1)$$

where each function $f_i : \mathbb{R}^p \to \mathbb{R}$ is α-strongly convex and β-smooth, and represents the local cost held by agent i.

Problems of this type have received a lot of attention in the last years, see *e.g.*, (Yi et al., 2016). Yet, most of these works assume that the composition of the system remains the same during the whole process. We extend such analyses to *open multi-agent systems* where agents can join or leave the network at a time-scale comparable to that of the process. Those are motivated by the growing size of the systems that tends to slow down the process as compared to the time-scale of potential arrivals and departures of agents. More generally, systems naturally allowing agents to join and leave are becoming common, such as *e.g.* multi-vehicle systems or with the Plug and Play implementation. In the case of (1), it results in the system size n_t, the local cost functions f_i^t, and the local demands d_i^t becoming time-varying. As a consequence, the solution of (1), denoted $x^{*,t}$, changes with the time as well, preventing classical convergence.

In this work, we analyze the behavior and performance of the Random Coordinate Descent algorithm (RCD), introduced in (Necoara, 2013), in open systems. The RCD algorithm solves (1) by using pairwise interactions, where at each step two randomly selected agents follow each other's local gradient to update their estimates, which is more appropriate for open systems, in particular where the number of agents can be large as it results in a low computational complexity. The convergence of this algorithm in open systems in terms of the distance to the minimizer in expectation was studied in (Monnoyer de Galland et al., 2021). In this work, we aim at analyzing the RCD algorithm from another perspective, following the assumption that the cost function must be paid at each iteration, so that a natural choice is to evaluate the performance accumulated over the iterations. Hence, we use metrics related to those used in online optimization, which commonly studies optimization with dynamical cost functions using a similar approach. In particular, we compare

⋆ C. Monnoyer de Galland and R. Vizuete equally contributed to this work. C. Monnoyer de Galland is a FRIA fellow (F.R.S.-FNRS). This work is supported by the "RevealFlight" ARC at UCLouvain, by the *Incentive Grant for Scientific Research (MIS)* "Learning from Pairwise Data" of the F.R.S.-FNRS and in part by the Agence Nationale de la Recherche (ANR) via grant "Hybrid And Networked Dynamical sYstems" (HANDY), number ANR-18-CE40-0010.

the performance of the RCD algorithm with that of the *optimal solution* $x^{*,t}$ and of the *selfish strategy*, denoted $x^{s,t}$, which consists in the total absence of collaboration between the agents (*i.e.*, $x_i^{s,t} = d_i^t$ at all times) and which is expected to yield less good performance. A representation of the behavior of these strategies is presented in Fig. 1.

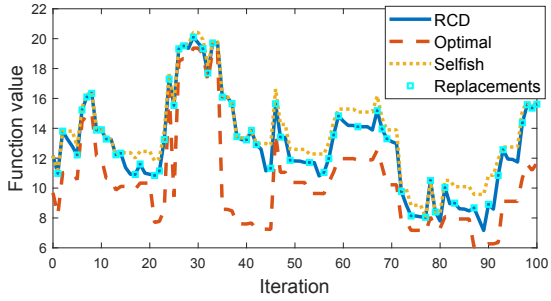


Fig. 1. Evolution of the function value f^t evaluated with the RCD algorithm, the optimal solution $x^{*,t}$, and the selfish strategy $x^{s,t}$, in a system subject to replacements of agents (*i.e.*, simultaneous departures and arrivals) happening as frequently as RCD steps.

2. PROBLEM FORMULATION

We consider the resource allocation problem defined in (1), where the local functions satisfy the following assumption.

Assumption 1. (Local cost function). At any time t, the local function f_i^t of any agent i is

- continuous differentiable;
- α-strongly convex: $f_i^t(x) - \frac{\alpha}{2}\|x\|^2$ is convex $\forall x$;
- β-smooth: $\|\nabla f_i^t(x) - \nabla f_i^t(y)\| \leq \beta\|x-y\|, \forall x, y$;
- satisfies $\arg\min_{x\in\mathbb{R}^d} f_i^t(x) = 0$ and $f_i^t(0) = 0$.

As a consequence of Assumption 1, the global cost function f^t is α-strongly convex and β-smooth as well at all time t. Assumption 1 also implies that the cost function of an agent i is zero only when $x_i = 0$, *i.e.*, when this agent does not contribute to satisfying the total demand.

In this preliminary work, we restrict our analysis to the specific setting described by the following assumptions.

Assumption 2. The local cost function of any agent i at any time t is one-dimensional: $f_i^t : \mathbb{R} \to \mathbb{R}$.

Assumption 3. (Homogeneous demand). The demand associated with any agent i at any time t is $d_i^t = 1$.

2.1 Random Coordinate Descent algorithm

We consider that agents interact through a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ is the set of agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. To solve (1), the agents perform the Random Coordinate Descent (RCD) algorithm introduced in (Necoara, 2013), such that at each iteration of the algorithm a pair of agents $(i, j) \in \mathcal{E}$ is uniformly randomly selected and i and j update their states according to

$$x_i^+ = x_i - \frac{1}{\beta}\left(f_i'(x_i) - f_j'(x_j)\right);$$
$$x_j^+ = x_j - \frac{1}{\beta}\left(f_j'(x_j) - f_i'(x_i)\right),$$

where we denote the derivative of the cost functions f_i with f_i' by simplicity because they are 1-dimensional. At the network level, the update rule can be expressed as:

$$x^+ = x - \frac{1}{\beta} Q^{ij} \nabla f(x), \tag{2}$$

where Q^{ij} is a $n \times n$ matrix filled with zeroes except for the following four entries:

$$\left[Q^{ij}\right]_{i,i} = \left[Q^{ij}\right]_{j,j} = \frac{1}{2}; \quad \left[Q^{ij}\right]_{i,j} = \left[Q^{ij}\right]_{j,i} = -\frac{1}{2}.$$

For this preliminary work we also make the following assumption, which means that pairwise interactions as defined in (2) can potentially take place between any pair of agents (i, j) in the systems.

Assumption 4. The graph $G = (\mathcal{V}, \mathcal{E})$ is complete.

2.2 Open system

We consider an *open* system, which is thus subject to arrivals and departures of agents. Those are modelled as follows, where we use respectively the subscripts *out* and *in* to refer to the leaving and joining agent:

(1) Departure: the agent *out* is uniformly randomly selected among the n agents in the system, and sends a last message to all its neighbours (*i.e.* the complete graph), such that for all $i \neq out$ there holds

$$x_i^+ = x_i + \frac{x_{out} - x_i}{n} = \left(1 - \frac{1}{n}\right) x_i + \frac{1}{n} x_{out}. \tag{3}$$

(2) Arrival: The agent *in* joins the system with its demand as initial value: $x_{in}^+ = d_{in} = 1$.

In this preliminary work, we restrict our analysis to open systems when only replacements take place (*i.e.* the simultaneous occurrence of a departure and of an arrival), so that the system size is fixed: for all t there holds $n_t = n$. We can then define the set of events possibly happening in the system as

$$\Xi := R \cup U = \left(\bigcup\nolimits_{i\in\mathcal{V}} R_i\right) \cup \left(\bigcup\nolimits_{(i,j)\in\mathcal{E}} U_{ij}\right), \tag{4}$$

where R_i denotes the replacement of agent i and U_{ij} a pairwise interaction between agents i and j leading to an update of their estimates according to (2).

We assume that two distinct events never occur simultaneously, so that we consider a discrete evolution of the time, where each time-step corresponds to an event $\xi \in \Xi$ happening in the system. Hence we can define the history of the process until a given time-step k as

$$\omega^k := \{(1, \xi_1), \dots, (k, \xi_k)\}, \tag{5}$$

where $\xi_i \in \Xi$ for all $i = 1, \dots, k$.

Assumption 5. The events ξ_i constituting a sequence ω^k happen independently of each other and of all past information, so that at each time-step either an update (*i.e.*, an event U) happens with fixed probability p or a replacement (*i.e.*, an event R) with fixed probability $1 - p$.

One can show that the definitions of arrivals, departures and of the RCD algorithm guarantee that at the occurrence of any event $\xi \in \Xi$, the estimate x^t remains feasible, *i.e.*, such that $x_i^t \geq 0$ for all i, and $\sum_{i=1}^N x_i^t = \sum_{i=1}^N d_i^t = n$, as long as the initial estimate x^0 is feasible as well.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

We can now reformulate original problem (1) in our particular setting as

$$\min_{x\in\mathbb{R}^n} f^t(x) = \sum_{i=1}^{n} f_i^t(x_i) \qquad \text{s.t.} \qquad \sum_{i=1}^{n} x_i = n. \tag{6}$$

Since the local cost functions satisfy Assumption 1, one can show that there exist $L_f, U_f \geq 0$ with $L_f \leq U_f$ such that $\forall t$ there holds

$$f^t(x^{*,t}) \in [L_f, U_f], \tag{7}$$

where $x^{*,t}$ is the solution of (6) at time t.

3. PERFORMANCE METRICS

Standard results on online optimization usually rely on the analysis of the so-called "*regret*", which quantifies the accumulation of the errors made by an algorithm as compared to the optimal solution over a finite number of time steps T. In this work, we follow a similar approach to analyze the performance of a given algorithm as compared to two possible strategies. Due to the definition of our problem, the conclusion is however expected to differ from classical online optimization. We consider the three following strategies for the study of the system:

- **Perfect collaboration**: we assume that at each time instant t the agents know the optimal solution denoted by $x^{*,t}$, which solves (6);
- **Selfish players**: we assume that the players do not collaborate to minimize f^t, and they operate at their individual desired point so that $x^{s,t} = \mathbb{1}_n$ at all t;
- **Decided strategy**: this case corresponds to the action given by a specific algorithm used to solve the problem, resulting in the estimate x^t at time t.

From the definitions, it is clear that for a well-designed algorithm it is expected that:

$$\sum_{t=1}^{T} f^t(x^{*,t}) \leq \sum_{t=1}^{T} f^t(x^t) \leq \sum_{t=1}^{T} f^t(x^{s,t}).$$

However, due to the replacements, there could be time instants where the selfish strategy behaves better than the decided strategy. An illustration of the evolution of these strategies is given in Fig. 1, where the decided strategy is the RCD algorithm defined in the previous section.

Based on these strategies and motivated by the techniques used on online optimization, we define these three quantities:

- **Dynamical Regret**:

$$\mathrm{Reg}_T := \sum\nolimits_{t=1}^{T} \left(f^t(x^t) - f^t(x^{*,t})\right); \tag{8}$$

- **Benefit**:

$$\mathrm{Ben}_T := \sum\nolimits_{t=1}^{T} \left(f^t(x^{s,t}) - f^t(x^t)\right); \tag{9}$$

- **Potential Benefit**:

$$\mathrm{Pot}_T := \sum\nolimits_{t=1}^{T} \left(f^t(x^{s,t}) - f^t(x^{*,t})\right). \tag{10}$$

The "*dynamical regret*" corresponds to the accumulation of the instantaneous errors obtained from using the decided strategy with respect to the optimal solution $x^{*,t}$. Similarly, the "*benefit*" quantifies the accumulated gain obtained from using the decided strategy instead of the selfish strategy $x^{s,t}$. Finally, the "*potential benefit*" represents the accumulated advantage of the optimal strategy with respect to the selfish one, and links the three quantities together as $\mathrm{Pot}_T = \mathrm{Ben}_T + \mathrm{Reg}_T$. Our goal is to analyze the evolution of all these quantities in expectation.

4. PRELIMINARY RESULTS

From Assumption 5, and following an approach similar to that used in (Monnoyer de Galland et al., 2021), we can study the evolution of the quantities defined in (8) – (10) in expectation by analyzing their evolution with each type of event independently (in our case, single replacements of agents and RCD iterations).

Observe that both $f^t(x^{*,t})$ and $f^t(x^{s,t})$ are only impacted by replacements, whereas $f^t(x^t)$ changes with replacements on the one hand, and on the other hand with RCD updates accordingly with the results presented in (Necoara, 2013). Hence, by properly analyzing the effect of replacements on the quantities of interest, we can show that the expected dynamical regret, the benefit and the potential benefit all grow in $O(T)$, so that on average over T they converge to a constant. In particular, we show the following result.

Theorem 1. In the setting described in Section 2, there holds

$$\lim_{T\to\infty} \frac{\mathbb{E}\mathrm{Pot}_T}{T} \leq \frac{\alpha}{2}(\kappa - 1)n; \tag{11}$$

$$\lim_{T\to\infty} \frac{\mathbb{E}\mathrm{Reg}_T}{T} \leq \rho_R \theta \kappa (n-1), \tag{12}$$

for some $\theta \geq 0$ which depends on β and α, and with $\rho_R = \frac{1-p}{p}$.

The detailed proof of Theorem 1 is omitted here, and will be presented elsewhere. Nevertheless, we provide a bit of insights about it below.

The first result of Theorem 1 is obtained from studying the largest possible improvement of using the optimal solution as compared to the selfish strategy at each time-step. In particular, using strong convexity and smoothness, one can show that $f^t(x^{s,t}) - f^t(x^{*,t}) \leq \frac{\beta}{2}\|\mathbb{1}\|^2 - \frac{\alpha}{2}\|x^{*,t}\|^2$, which ultimately yields (11). The second result of Theorem 1 relies on the separated analysis of $f^{t+1}(x^{t+1}) - f^t(x^t)$ and $f^{t+1}(x^{*,t+1}) - f^t(x^{*,t})$, which allow building the evolution of $f^t(x^t) - f^t(x^{*,t})$ as a recurrence on these quantities over the iterations. One can then show that the sum of the latter over the iterations is bounded using (7), thus avoiding the introduction of an additive term related to the optimal value of the function at each time-step. Additional computations then ultimately lead to (12).

Consistently with the definition of the problem and of the quantities of interest, the results above depend on the condition number κ and on the system size n. Moreover, only the regret, which depends on RCD updates, depends on $\rho_R = \frac{1-p}{p}$, *i.e.*, the expected number of replacements between two RCD updates, and thus illustrates how the openness of the system interferes in the behavior of the algorithm. Finally, the qualitative behavior of the bounds matches that of the simulated results obtained with a system of 5 agents subject to replacements, as illustrated in Fig. 2. The derivation of a proper upper bound on the

expected benefit, parallel to (11) and (12) is the next step in this analysis. Furthermore, deriving the corresponding lower bounds and showing that they grow linearly with T as well would be the natural prosecution of these preliminary results, as they would validate the behavior illustrated in Fig. 2 and show that it is not possible to obtain a sublinear behavior.

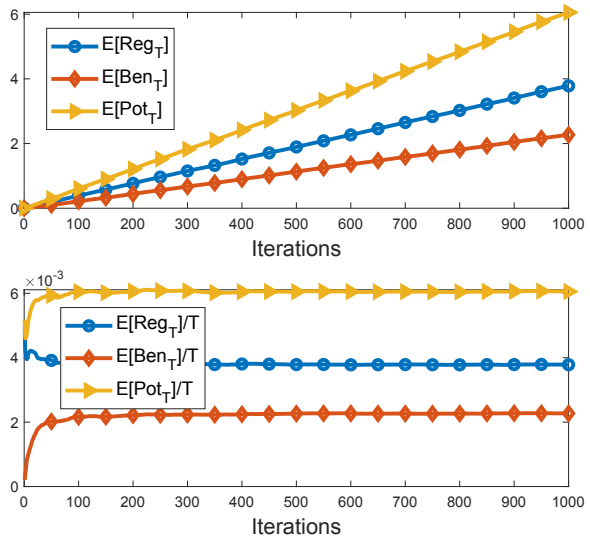


Fig. 2. Evolution of $\mathbb{E}\mathrm{Reg}_T$, $\mathbb{E}\mathrm{Ben}_T$ and $\mathbb{E}\mathrm{Pot}_T$ (top) and of the corresponding quantities averaged with respect to the time (bottom) simulated in a system of 5 agents subject to replacements as frequent as RCD steps.

5. CONCLUSION

We analyzed the performance and behavior of the Random Coordinate Descent algorithm (RCD) for solving the optimal resource allocation problem in an open system subject to replacements of agents, resulting in variation of the total cost function and of the total amount of resource to be allocated. We considered a simple preliminary setting where the budget is homogeneous and where the graph is complete, and used tools inspired from online optimization to show that it is not possible to achieve convergence to the optimal solution with the RCD algorithm in expectation in open system, but that the error is expected to remain reasonable.

We have derived upper bounds on the evolution of the regret and the potential benefit in expectation. A natural continuation of this work is thus the derivation of the corresponding upper bound for the benefit, and of lower bounds for this quantities in order to validate the observed behavior. More generally, our bounds could be extended to more general settings, and their tightness can be improved to match more accurately the actual performance of the algorithm. Moreover, since our approach is based on the analysis of the effect of arrivals and departures of agents combined into replacements, the next step of this study is to generalize it to the case where the system size changes with the time, *i.e.*, where arrivals and departures are decoupled.

REFERENCES

Dominguez-Garcia, A.D., Cady, S.T., and Hadjicostis, C.N. (2012). Decentralized optimal dispatch of distributed energy resources. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 3688–3693. IEEE.

Monnoyer de Galland, C., Vizuete, R., Hendrickx, J.M., Frasca, P., and Panteley, E. (2021). Random coordinate descent algorithm for open multi-agent systems with complete topology and homogeneous agents. In *2021 IEEE 60th Conference on Decision and Control (CDC)*, 1701–1708. IEEE.

Necoara, I. (2013). Random coordinate descent algorithms for multi-agent convex optimization over networks. *IEEE Transactions on Automatic Control*.

Teixeira, A., Araújo, J., Sandberg, H., and Johansson, K.H. (2013). Distributed actuator reconfiguration in networked control systems. *IFAC Proceedings Volumes*, 46(27), 61–68.

Yi, P., Hong, Y., and Liu, F. (2016). Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems. *Automatica*, 74, 259–269.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Data-driven Approximation of the Koopman Generator on Rich Approximation Spaces

Feliks Nüske*

**Max-Planck-Institute for Dynamics of Complex Technical Systems, Magdeburg, 39106 Germany (e-mail: nueske@mpi-magdeburg.mpg.de)*

Abstract: In the context of Koopman operator based analysis of dynamical systems, the generator of the Koopman semigroup is of central importance. Models for the Koopman generator can be used, among others, for system identification, coarse graining, and control of the system at hand. A critical modeling choice is the subspace or dictionary used for Koopman estimation. In this talk, I will present recent advances allowing for the approximation of the generator on reproducing kernel Hilbert spaces (RKHS), and on tensor-structured subspaces by means of low-rank representations. Both approaches allow modelers to employ high-dimensional, or even infinite-dimensional approximation spaces, while controlling the computational effort at the same time. In both cases, I will discuss the algorithmic realization and computational complexity in detail. I will also discuss recent results on estimating the finite-data estimation error for Koopman generator models.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Input-to-state stability for unbounded bilinear feedback systems ⋆

René Hosfeld * **Birgit Jacob** ** **Felix Schwenninger** ***
Marius Tucsnak ****

* *University of Wuppertal, 42119 Wuppertal, and University of Hamburg, 20146 Hamburg, Germany (e-mail: hosfeld@uni-wuppertal.de).*
** *University of Wuppertal, 42119 Wuppertal, Germany (e-mail: bjacob@uni-wuppertal.de).*
*** *University of Twente, 7522NH Enschede, The Netherlands and University of Hamburg, 20146 Hamburg, Germany (e-mail: f.l.schwenninger@utwente.nl).*
**** *University of Bordeaux, 33405 Talence, France (e-mail: marius.tucsnak@u-bordeaux.fr).*

Abstract: We study input-to-state stability (ISS) of systems with a linear control and a bilinear feedback term, depending on the state trajectory itself and the output of the system. Both, the control and the bilinear feedback enter the system through possibly unbounded operators. Further, the observation operator, associated to the output, is also considered to be unbounded. We derive sufficient conditions for a global in time well-posedness result for small initial data as well as sufficient conditions for an ISS-estimate. This extends recent investigations on bilinear systems, where a second control was considered instead of an output. The developed results are applied to a Burgers equation.

Keywords: Input-to-state stability, well-posed distributed parameter systems, semigroup and operator theory, stability of nonlinear systems, robust control.

1. INTRODUCTION

The concept of *input-to-state stability* (ISS) as introduced in Sontag (1989) unifies both asymptotic stability with respect to initial values and robustness with respect to external inputs. More specific, if we consider a system Σ as a mapping which maps initial values $z_0 \in X$ and inputs $u : [0,\infty) \to U$ to the (unique) state trajectory $z : [0,T) \to X$, defined on its maximal existence interval, then, Σ is said to be ISS if for all initial values $z_0 \in X$ and inputs $u \in \mathrm{L}^\infty_{\mathrm{loc}}(0,\infty;U)$ there exists a unique global (i.e $T = \infty$) state trajectory and for all $t \in [0,\infty)$,

$$\|z(t)\|_X \leq \beta(\|z_0\|,t) + \gamma(\|u\|_{\mathrm{L}^\infty(0,t;U)}),$$

for some functions $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ where the Lyapunov classes $\mathcal{K}$ and $\mathcal{KL}$ are given by

$$\mathcal{K} = \left\{\gamma \in C(\mathbb{R}_0^+;\mathbb{R}_0^+) \,\middle|\, \gamma \text{ is strictly increasing, } \gamma(0) = 0\right\}$$

$$\mathcal{KL} = \left\{ \beta \in C(\mathbb{R}_0^+ \times \mathbb{R}_0^+;\mathbb{R}_0^+) \,\middle|\, \begin{array}{c} \beta(\cdot,t) \in \mathcal{K} \ \forall t \geq 0 \text{ and} \\ \beta(s,\cdot) \text{ is strictly} \\ \text{decreasing to } 0 \ \forall s > 0 \end{array} \right\}.$$

The Banach spaces X and U with norms $\|\cdot\|_X$ and $\|\cdot\|_U$ are called the *state space* and *input space* of Σ.
Often one would like to consider a larger class of inputs than $\mathrm{L}^\infty_{\mathrm{loc}}$. For this reason we consider the following adoption which is referred to as L^2-ISS

$$\|z(t)\|_X \leq \beta(\|z_0\|,t) + \gamma(\|u\|_{\mathrm{L}^2(0,t;U)}),$$

with β and γ as before.

In this work abstract we investigate L^2-ISS of unbounded bilinear feedback systems

$$\begin{cases} \dot{z}(t) = Az(t) + B_1u_1(t) + B_2N(z(t),y(t)), \\ y(t) = Cz(t), \end{cases} \tag{1}$$

where A is the generator of a C_0-semigroup and B_i, $i = 1,2$, and C are possibly unbounded operators. The precise setting is introduced in Section 2.
For the simple case of linear systems

$$\dot{z}(t) = Az(t) + Bu(t)$$

where $B \in \mathcal{L}(U,X)$, i.e. B is bounded from U to X, L^2-ISS is equivalent to uniform exponential stability of the semigroup. For unbounded B, which appears naturally when considering boundary control, L^2-ISS becomes non-trivial. Indeed, it is related to L^2-admissibility of the control operator B, see e.g. Jacob et al. (2018).

We refer to Karafyllis and Krstic (2019); Mironchenko and Prieur (2020); Schwenninger (2020) for an overview on ISS for infinite-dimensional systems and to Sontag (2008) for a survey on ISS for ODEs. Recent results for infinite-dimensional systems can be found in Dashkovskiy and Mironchenko (2013a,b); Guiver et al. (2019); Jayawardhana et al. (2008); Mironchenko and Wirth (2018) and in Jacob et al. (2019); Karafyllis and Krstic (2017); Mazenc and Prieur (2011); Mironchenko and Ito (2015); Mironchenko et al. (2019); Zheng and Zhu (2018) for (semi-)linear systems with a slight focus on the parabolic case.

⋆ The first three authors have been supported by the German Research Foundation (DFG) via the joint grant JA 735/18-1 / SCHW 2022/2-1.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The results from Mironchenko and Ito (2016) on ISS for bounded bilinear control systems has been generalized in Hosfeld et al. (2022) to unbounded bilinear control systems of the form

$$\dot{z}(t) = Az(t) + B_1u_1(t) + B_2N(z(t), u_2(t))$$

with unbounded operators B_1, B_2 and suitable boundedness and Lipschitz assumptions on the non-linearity N. The question of ISS becomes more delicate as soon as a feedback law $u_2 = y = Cz$ is present which turns the above system into (1). Even the question of well-posedness of such systems is far from being evident. A local in time well-posedness result for (1) can be found in Tucsnak and Weiss (2014). This is insufficient for our considerations since global in time well-posedness is a necessity for ISS.

This work is structured in the following way. In Section 2 we formalize the setting of a bilinear feedback system. Afterwards we recall the concept of well-posedness of linear systems in order to state the solution concept of the bilinear feedback system. We close the section with an existence and uniqueness result for the solution of the bilinear feedback system with small input data. In Section 3 an ISS result on the bilinear feedback system for certain parabolic and hyperbolic systems is given which we apply in Section 4 to a Burgers equation.

2. A BILINEAR FEEDBACK SYSTEM

Let X, U_1, U_2, Y be Banach spaces. We now consider the unbounded bilinear feedback system of the form

$$\begin{cases} \dot{z}(t) &= Az(t) + B_1u_1 + B_2N(z(t), y(t)),\ t \geq 0, \\ z(0) &= z_0, \\ y(t) &= Cz(t),\ t \geq 0, \end{cases} \tag{Σ^N}$$

where

- A generates a C_0-semigroup $(T(t))_{t\geq 0}$ on X,
- $B_1 \in \mathcal{L}(U_1, X_{-1})$, $B_2 \in \mathcal{L}(U_2, X_{-1})$ are the (unbounded) *control operators*,
- $C \in \mathcal{L}(X_1, Y)$ is the (unbounded) *observation operator*,
- $Y \subseteq X$ with continuous embedding and C has an extension $C \in \mathcal{L}(X)$,
- $N : X \times Y \to U_2$ is a continuous bilinear mapping which satisfies for some $K > 0$, $p \in (0, 1)$ and all $x \in X$, $y \in Y$

$$\|N(z, y)\|_{U_2} \leq K\|z\|_X\|y\|_X^{1-p}\|y\|_Y^p. \tag{2}$$

The spaces X, U_1, U_2 and Y are referred as the *state space* (X), the *input spaces* (U_1, U_2) and the *output space* (Y). By X_1 and X_{-1} we denote the standard inter- and extrapolation spaces associated to the generator A. That is, X_1 is $D(A)$, the domain of A, equipped with the graph norm and X_{-1} is the completion of X with respect to the norm

$$\|x\|_{X_{-1}} = \|(\beta - A)^{-1}x\|_X$$

for some $\beta \in \rho(A)$, the resolvent set of A. We have the continuous and dense embeddings $X_1 \hookrightarrow X \hookrightarrow X_{-1}$. Furthermore, $(T(t))_{t\geq 0}$ extends uniquely to a C_0-semigroup $(T_{-1}(t))_{t\geq 0}$ on X_{-1} whose generator, denoted by A_{-1}, is an extension of A with domain $D(A_{-1}) = X$. By $\omega_0((T(t))_{t\geq 0})$ we denote the the growth bound of the semigroup $(T(t))_{t\geq 0}$. The semigroup $(T(t))_{t\geq 0}$ is exponentially stable if $\omega_0((T(t))_{t\geq 0}) < 0$, i.e. there exist $M, \omega_0 > 0$ such that

$$\|T(t)\|_X \leq M\,\mathrm{e}^{-\omega_0 t} \quad \text{for all } t \geq 0. \tag{3}$$

Next we recall the definition and standard facts on well-posedness (in the L^2-sense, see e.g. Tucsnak and Weiss (2014); Weiss (1989a,b); Curtain and Weiss (1989)) of the to Σ^N associated linear systems

$$\begin{cases} \dot{x}(t) &= Ax(t) + B_1u_1(t) + B_2u_2(t),\ t \geq 0, \\ x(0) &= x_0, \\ y(t) &= Cx(t),\ t \geq 0 \end{cases} \tag{Σ_{lin}}$$

with spaces and operators as before. For simplicity we understand $\mathrm{L}^2(0, t; U)$ as a subspace of $\mathrm{L}^2(0, \infty; U)$ with the usual conventions of truncating and extending functions by zero. The linear system Σ_{lin} is *well-posed*, if

(i) A generates a C_0-semigroup $(T(t))_{t\geq 0}$ on X;

(ii) B_i, $i = 1, 2$, are L^2-*admissible control operators*, i.e. for some (and hence for all) $t > 0$ it holds that the *input maps* $\Phi_t^i \in \mathcal{L}(\mathrm{L}^2(0, t; U_i), X)$, where

$$\Phi_t^i u_i := \int_0^t T_{-1}(t - s)B_iu_i(s)\,\mathrm{d}s, \quad u_i \in \mathrm{L}^2(0, t; U_i);$$

(iii) C is an L^2-*admissible observation operator*, i.e. for some (and hence for all) $t > 0$ it holds that the *output maps* $\Psi_t \in \mathcal{L}(X, \mathrm{L}^2(0, t; Y))$, where Ψ_t is the extension of

$$\Psi_t x = CT(\cdot)x, \quad x \in D(A);$$

(iv) Some (and hence any) function $G_i : \mathbb{C}_{\omega_0((T(t))_{t\geq 0})} \to \mathcal{L}(U_i, Y)$, $i = 1, 2$, satisfying

$$G_i(\alpha) - G_i(\beta) = C[(\alpha - A_{-1})^{-1} - (\beta - A_{-1})^{-1}]B_i$$

is bounded on some right half-plane $\mathbb{C}_\gamma = \{\lambda \in \mathbb{C} \mid \mathrm{Re}\,\lambda > \gamma\}$.

The functions G_i, $i = 1, 2$ are called the *transfer functions* associated to the triples (A, B_i, C), $i = 1, 2$ and they are unique up to a constant operator. Given such a transfer function, we can define the so called *input output maps* $\mathbb{F}_t^i : \mathrm{L}^2(0, t; U_i) \to \mathrm{L}^2(0, t; Y)$, which are given by

$$\begin{aligned}(\mathbb{F}_t^i u_i)(\cdot) = C\left[\int_0^\cdot T_{-1}(\cdot - s)B_iu_i(s)\,\mathrm{d}s - (\alpha - A)^{-1}B_iu_i(\cdot)\right] \\ + G_i(\alpha)u_i(\cdot),\end{aligned}$$

where $u_i \in W_0^{1,2}(0, t; U_i)$ (dense in $\mathrm{L}^2(0, t; U_i)$) and $\mathrm{Re}\,\alpha$ large enough. To see that $\mathbb{F}_t^i$ is well-defined we refer to Curtain and Weiss (1989).

The functions $z \in C([0, \infty); X)$ and $y \in \mathrm{L}^2_{\mathrm{loc}}(0, \infty; Y)$ are called the *solution* (or *state trajectory*) and the *output function* (or *output*) of Σ_{lin} with input data $z_0 \in X$ and $u_i \in \mathrm{L}^2(0, \infty; U_i)$, $i = 1, 2$, if they satisfy

$$\begin{aligned} z(t) &= T(t)z_0 + \Phi_t^1u_1 + \Phi_t^2u_2, \\ y_{|_{[0,t]}} &= \Psi_t z_0 + \mathbb{F}_t^1u_1 + \mathbb{F}_t^2u_2. \end{aligned}$$

In particular z satisfies the integrated version of the differential equation in Σ_{lin}. From the above formula it follows immediately that a well-posed linear system admits a unique solution z and and output y for all $z_0 \in X$ and $u_i \in \mathrm{L}^2(0, \infty; U_i)$ and for $t \geq 0$ there exist positive constants $k_{1,t}$ and $k_{2,t}$ such that

$$\begin{aligned} \|z(t)\|_X &\leq k_{1,t}(\|z_0\|_X + \|u_1\|_{\mathrm{L}^2(0,t;U_1)} + \|u_2\|_{\mathrm{L}^2(0,t;U_2)}), \\ \|y\|_{\mathrm{L}^2(0,t;Y)} &\leq k_{2,t}(\|z_0\|_X + \|u_1\|_{\mathrm{L}^2(0,t;U_1)} + \|u_2\|_{\mathrm{L}^2(0,t;U_2)}). \end{aligned} \tag{4}$$

If A generates an exponential stable semigroup, then $k_{1,t}$ and $k_{2,t}$ can be chosen independent of t, see e.g. Weiss (1989b,a); Curtain and Weiss (1989).

Definition 1. Let Σ_{lin} be well-posed. The functions $z \in C([0,\infty);X)$ and $y \in \mathrm{L}^2_{\text{loc}}(0,\infty;Y)$ are called the *solution* and *output* of Σ^N with input data $z_0 \in X$ and $u_1 \in \mathrm{L}^2(0,\infty;U_1)$ if they satisfy

$$z(t) = T(t)z_0 + \Phi^1_t u_1 + \Phi^2_t N(z,y),$$
$$y_{|_{[0,t]}} = \Psi_t z_0 + \mathbb{F}^1_t u_1 + \mathbb{F}^2_t N(z,y).$$

Theorem 2. Let A be the generator of the exponentially stable semigroup $(T(t))_{t\geq 0}$ and choose $M, \omega_0 > 0$ according to (3). If Σ_{lin} is well-posed then for every $\omega \in (0,\omega_0)$ there exists a constant $\varepsilon > 0$ such that for all input data $z_0 \in X$ and $u_1 \in \mathrm{L}^2(0,\infty;U_1)$ with

$$\|z_0\|_X + \|\mathrm{e}^{\omega\cdot}\, u_1\|_{\mathrm{L}^2(0,\infty;U_1)} \leq \varepsilon \tag{5}$$

it holds that Σ^N admits a unique solution $z \in C([0,\infty);X)$ and an output $y \in \mathrm{L}^2(0,\infty;Y)$. Furthermore, there exists a constant $k_1 > 0$ such that

$$\|z(t)\|_X \leq 2k_1\varepsilon\, \mathrm{e}^{-\omega t}\,. \tag{6}$$

3. ISS-ESTIMATES FOR A SPECIAL CLASS OF UNBOUNDED BILINEAR FEEDBACK SYSTEMS

Throughout this section we assume X, U_1, U_2, Y to be Hilbert spaces.

Definition 3. System Σ^N is called L^2-ISS with respect to a non-empty subset $\mathcal{S} \subseteq X \times \mathrm{L}^2(0,\infty;U_1)$ if Σ^N admits for all input data $(z_0,u_1) \in \mathcal{S}$ a unique solution $z \in C([0,\infty);X)$ which satisfies

$$\|z(t)\|_X \leq \beta(\|z_0\|_X, t) + \gamma(\|u_1\|_{\mathrm{L}^2(0,t;U_1)})$$

for all $t \geq 0$ and some functions $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$.
If $\mathcal{S} = \{(z_0,u_1) \in X \times \mathrm{L}^2(0,\infty;U_1) \mid z_0, u_1 \text{ satisfy (5)}\}$ for some $\varepsilon > 0$ then Σ^N is called L^2-ISS for small input data (independent of $\varepsilon > 0$).

Recall that an operator A on X is called strictly dissipative if there exists a constant $w_A < 0$ such that

$$\operatorname{Re}\langle Az, z\rangle_X \leq w_A\|z\|^2_X \quad \text{for all } x \in D(A). \tag{7}$$

Additionally, if $\langle Az,z\rangle_X \in \mathbb{R}$ for all $z \in D(A)$, then A is called strictly negative.
We consider the following two assumptions:

Assumption 1. The operator A is selfadjoint and strictly negative, $B_i \in \mathcal{L}(U_i, X_{-\frac{1}{2}})$ and $C \in \mathcal{L}(X_{\frac{1}{2}}, Y)$, where the spaces $X_{\frac{1}{2}}$ and $X_{-\frac{1}{2}}$ are introduced below.

Assumption 2. The operator A is of the form $A = A_0 + K$, where A_0 is skew-adjoint and $K \in \mathcal{L}(X)$ is strictly dissipative, $B_i \in \mathcal{L}(U_i, X)$, $i = 1,2$ and $C \in \mathcal{L}(X,Y)$.

Remark 4. Under Assumption 2, A is the strictly dissipative generator of an exponentially stable semigroup and one could choose $w_A = w_K$, where w_A and w_K are the (negative) constants according to (7).

If A is a selfadjoint and strictly dissipative operator on X then it follows from Engel and Nagel (2000), Chapter II, Corollary 4.7 & Chapter IV, Corollary 3.12 and the fact that $\sigma(A) \subset (-\infty, w_A]$ that A is the generator of an exponentially stable analytic semigroup on X. One can define (c.f. Tucsnak and Weiss (2009), Section 3.4) the spaces $X_{\frac{1}{2}}$ as the completion of $D(A)$ with respect to the norm

$$\|z\|_{X_{\frac{1}{2}}} = \sqrt{\langle -Az, z\rangle_X}, \quad z \in D(A)$$

and $X_{-\frac{1}{2}}$ as its dual space with respect to the pivot space X. We have the continuous and dense embeddings

$$X_1 \hookrightarrow X_{\frac{1}{2}} \hookrightarrow X \hookrightarrow X_{-\frac{1}{2}} \hookrightarrow X_{-1}.$$

From the concept of dual spaces with respect to pivot space and the fact that A is selfadjoint we obtain a natural continuous dual pairing $\langle\cdot,\cdot\rangle_{X_{-\frac{1}{2}},X_{\frac{1}{2}}} : X_{-\frac{1}{2}} \times X_{\frac{1}{2}} \to \mathbb{C}$ which simplifies for $z \in X$ and $y \in X_{\frac{1}{2}}$ to

$$\langle z,y\rangle_{X_{-\frac{1}{2}},X_{\frac{1}{2}}} = \langle z,y\rangle_X\,.$$

Further, it is easy to see that $A_{-1} : X_{\frac{1}{2}} \to X_{-\frac{1}{2}}$ is isometric. This allows to extend (7) to

$$\langle A_{-1}z, z\rangle_{X_{-\frac{1}{2}},X_{\frac{1}{2}}} = -\|z\|_{X_{\frac{1}{2}}} \leq w_A\|z\|^2_X, \tag{8}$$

for every $z \in X_{\frac{1}{2}}$.

Remark 5. In order to obtain such a dual pairing it is crucial to have $D(A) = D(A^*)$.

Theorem 6. If Assumption 1 or Assumption 2 holds and for some $\delta > 0$ there exist $m_1, m_2 \in \mathbb{R}$ such that

$$1 - m_1 > 0, \quad (1-m_1)w_A + m_2 < 0$$

and

$$\operatorname{Re}\langle N(z,Cz), B_2^*z\rangle_{U_2} \leq -m_1 \operatorname{Re}\langle Az,z\rangle_X + m_2\|z\|^2_X \tag{9}$$

for all $z \in D(A)$ with $\|z\|_X \leq \delta$, then there exists $\varepsilon > 0$ such that Σ^N admits for all small input data $z_0 \in X$ and $u_1 \in \mathrm{L}^2(0,\infty;U_1)$ in the sense of (5) a unique solution z which satisfies

$$\|z(t)\|_X \leq \mathrm{e}^{-\omega t}\|z_0\|_X + c\|u_1\|_{\mathrm{L}^2(0,t;U_1)} \tag{10}$$

for all $t \geq 0$ and some constants $\omega, c > 0$. In particular Σ^N is L^2-ISS for small input data.

4. EXAMPLE: BURGERS EQUATION

Consider the following controlled version of the Burgers equation

$$\begin{cases} \dot z(t,x) = z_{xx}(t,x) - zz_x(t,x) + u_1(t), & t > 0,\ x \in (0,1) \\ z(t,0) = z(t,1) = 0, & t \geq 0, \\ z(0,x) = z_0(x), & x \in [0,1], \\ y(t,x) = z(t,x), & t \geq 0,\ x \in [0,1]. \end{cases}$$

Let the state, input and output spaces be given as in Tucsnak and Weiss (2014), i.e.

$$X = \mathrm{H}^1_0(0,1),$$
$$U_1 = U_2 = \mathrm{L}^2(0,1),$$
$$Y = \mathrm{H}^2(0,1) \cap \mathrm{H}^1_0(0,1),$$

where all spaces are assumed to be real valued. We equip $\mathrm{H}^1_0(0,1)$ with the norm

$$\|z\|_{\mathrm{H}^1_0} = \|z_x\|_{\mathrm{L}^2}.$$

which is equivalent to the standard norm on $\mathrm{H}^1_0(0,1)$ by the Poincaré inequality.
Let the operator A on X be defined by

$$A\varphi = \varphi_{xx}, \quad D(A) = \{\varphi \in \mathrm{H}^3(0,1) \mid \varphi, \varphi_{xx} \in \mathrm{H}^1_0(0,1)\}.$$

It is known that A is a selfadjoint and strictly dissipative and hence the generator of an exponentially stable analytic semigroup.

Note that (see Tucsnak and Weiss (2014) and the references therein)

$$X_{\frac{1}{2}} = \mathrm{H}^2(0,1) \cap \mathrm{H}_0^1(0,1) \quad \text{and} \quad X_{-\frac{1}{2}} = \mathrm{L}^2(0,1).$$

Further, we consider the operators $B_i \in \mathcal{L}(U_i, X_{-\frac{1}{2}})$, $i = 1, 2$ and $C \in \mathcal{L}(X_{\frac{1}{2}}, Y)$ to be the identity on the respective spaces. The bilinear feedback law is defined by

$$N : X \times Y \to U_2, \quad N(z, y) = -zy_x.$$

The validity of (2) for any $p \in (0, 1)$ is not difficult to check and can also be found in Tucsnak and Weiss (2014).

Theorem 7. The Burgers equation (4), with the above spaces and operators, is a bilinear feedback system of the form Σ^N which is L^2-ISS for small input data $z_0 \in \mathrm{H}_0^1(0,1)$ and $u_1 \in \mathrm{L}^2(0, \infty; \mathrm{L}^2(0,1))$.

REFERENCES

Curtain, R.F. and Weiss, G. (1989). Well posedness of triples of operators (in the sense of linear systems theory). In *Control and estimation of distributed parameter systems (Vorau, 1988)*, volume 91 of *Internat. Ser. Numer. Math.*, 41–59. Birkhäuser, Basel.

Dashkovskiy, S. and Mironchenko, A. (2013a). Input-to-state stability of infinite-dimensional control systems. *Math. Control Signals Systems*, 25(1), 1–35. doi:10.1007/s00498-012-0090-2.

Dashkovskiy, S. and Mironchenko, A. (2013b). Input-to-state stability of nonlinear impulsive systems. *SIAM J. Control Optim.*, 51(3), 1962–1987. doi:10.1137/120881993.

Engel, K.J. and Nagel, R. (2000). *One-parameter semigroups for linear evolution equations*, volume 194 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

Guiver, C., Logemann, H., and Opmeer, M.R. (2019). Infinite-Dimensional Lur'e Systems: Input-To-State Stability and Convergence Properties. *SIAM J. Control Optim.*, 57(1), 334–365. doi:10.1137/17M1150426.

Hosfeld, R., Jacob, B., and Schwenninger, F. (2022). Integral input-to-state stability of unbounded bilinear control systems. *Mathematics of Control, Signals, and Systems*. doi:10.1007/s00498-021-00308-9.

Jacob, B., Nabiullin, R., Partington, J., and Schwenninger, F. (2018). Infinite-dimensional input-to-state stability and Orlicz spaces. *SIAM J. Control Optim.*, 56(2), 868–889.

Jacob, B., Schwenninger, F., and Zwart, H. (2019). On continuity of solutions for parabolic control systems and input-to-state stability. *J. Differential Equations*, 266(10), 6284–6306. doi:10.1016/j.jde.2018.11.004.

Jayawardhana, B., Logemann, H., and Ryan, E. (2008). Infinite-dimensional feedback systems: the circle criterion and input-to-state stability. *Commun. Inf. Syst.*, 8(4), 413–444.

Karafyllis, I. and Krstic, M. (2017). ISS in different norms for 1-D parabolic PDEs with boundary disturbances. *SIAM J. Control Optim.*, 55(3), 1716–1751.

Karafyllis, I. and Krstic, M. (2019). *Input-to-state stability for PDEs*. Communications and Control Engineering Series. Springer, Cham. doi:10.1007/978-3-319-91011-6.

Mazenc, F. and Prieur, C. (2011). Strict Lyapunov functions for semilinear parabolic partial differential equations. *Math. Control Relat. Fields*, 1(2), 231–250. doi:10.3934/mcrf.2011.1.231.

Mironchenko, A. and Ito, H. (2015). Construction of Lyapunov Functions for Interconnected Parabolic Systems: An iISS Approach. *SIAM J. Control Optim.*, 53(6), 3364–3382. doi:10.1137/14097269X.

Mironchenko, A. and Ito, H. (2016). Characterizations of integral input-to-state stability for bilinear systems in infinite dimensions. *Math. Control Relat. Fields*, 6(3), 447–466.

Mironchenko, A., Karafyllis, I., and Krstic, M. (2019). Monotonicity Methods for Input-to-State Stability of Nonlinear Parabolic PDEs with Boundary Disturbances. *SIAM J. Control Optim.*, 57(1), 510–532. doi:10.1137/17M1161877.

Mironchenko, A. and Prieur, C. (2020). Input-to-State Stability of Infinite-Dimensional Systems: Recent Results and Open Questions. *SIAM Rev.*, 62(3), 529–614. doi:10.1137/19M1291248.

Mironchenko, A. and Wirth, F. (2018). Characterizations of input-to-state stability for infinite-dimensional systems. *IEEE Trans. Automat. Control*, 63(6), 1602–1617. doi:10.1109/tac.2017.2756341.

Schwenninger, F. (2020). Input-to-state stability for parabolic boundary control: Linear and semi-linear systems. In J. Kerner, L. Laasri, and D. Mugnolo (eds.), *Control Theory of Infinite-Dimensional Systems*, 83–116. Birkhäuser, Cham.

Sontag, E. (1989). Smooth stabilization implies coprime factorization. *IEEE Trans. Automat. Control*, 34(4), 435–443. doi:10.1109/9.28018.

Sontag, E. (2008). Input to state stability: basic concepts and results. In *Nonlinear and optimal control theory*, volume 1932 of *Lecture Notes in Math.*, 163–220. Springer Berlin.

Tucsnak, M. and Weiss, G. (2009). *Observation and Control for Operator Semigroups*. Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Verlag, Basel.

Tucsnak, M. and Weiss, G. (2014). Well-posed systems—the LTI case and beyond. *Automatica J. IFAC*, 50(7), 1757–1779. doi:10.1016/j.automatica.2014.04.016.

Weiss, G. (1989a). Admissibility of unbounded control operators. *SIAM J. Control Optim.*, 27(3), 527–545. doi:10.1137/0327028.

Weiss, G. (1989b). Admissible observation operators for linear semigroups. *Israel J. Math.*, 65(1), 17–43. doi:10.1007/BF02788172.

Zheng, J. and Zhu, G. (2018). Input-to-state stability with respect to boundary disturbances for a class of semi-linear parabolic equations. *Automatica J. IFAC*, 97, 271–277. doi:10.1016/j.automatica.2018.08.007.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A deterministic least squares approach for simultaneous input and state estimation ⋆

Grigorios Gakis * **Malcolm C. Smith** *

* *Department of Engineering, University of Cambridge, UK, (e-mail: gg402@cam.ac.uk and mcs@eng.cam.ac.uk).*

Abstract: This paper considers a deterministic estimation problem to find the input and state of a linear dynamical system which minimise a weighted integral squared error between the resulting output and the measured output. A completion of squares approach is used to find the unique optimum in terms of the solution of a Riccati differential equation. The optimal estimate is obtained from a two-stage procedure that is reminiscent of the Kalman filter. The first stage is an end-of-interval estimator for the finite horizon which may be solved in real time as the horizon length increases. The second stage computes the unique optimum over a fixed horizon by a backwards integration over the horizon. A related tracking problem is solved in an analogous manner. Making use of the solution to both the estimation and tracking problems a constrained estimation problem is solved which shows that the Riccati equation solution has a least squares interpretation that is analogous to the meaning of the covariance matrix in stochastic filtering. The paper shows that the estimation and tracking problems considered here include the Kalman filter and the linear quadratic regulator as special cases.

Keywords: Input and state estimation, least squares optimisation, Kalman filtering, optimal trajectory tracking.

1. INTRODUCTION

Our goal in this paper is to pose and solve a filtering/estimation problem for the simultaneous estimation of inputs and states in a continuous time linear finite dimensional dynamical system. The problem set-up is illustrated in Fig. 1. A model of the physical system is assumed to be available. The output of the dynamical system is the vector z of *all* variables that are measured (e.g. by means of sensors). The measurement of this output in an experiment is denoted by $\tilde{z}$. The filter should make use only of these measured outputs for the estimation and should produce the best estimate of the system variables treating the exogenous inputs and states on an equal footing. The meaning of "best" is to minimise a weighted integral squared error between output z and measured output $\tilde{z}$. The filtered signals $w_1(t), x_1(t)$ provide best estimates of w and x at a given time instant t based on measurements up to that time, while the estimates $\hat{w}(t), \hat{x}(t)$ provide the best estimates over a time interval $[0, T]$.

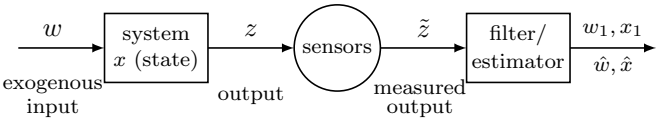


Fig. 1. Estimation problem for a dynamical system with state x, exogenous input w and output z which is measured.

Our solution to the problem of Fig. 1 is based on the method of completion of squares and builds on the work of Willems (2004) which gave a deterministic derivation of the filter of Kalman and Bucy (1961). The structure of the solution is reminiscent of the standard (causal) Kalman filter and the non-causal process of smoothing, though the solution to the present problem has a more general structure through placing the estimation of state and exogenous input on an equal footing, and without prior assumption on the nature of the exogenous input. An important further contribution of the paper is to provide a deterministic interpretation of the solutions of the resulting matrix Riccati differential equations (which correspond to the covariance matrices in the standard Kalman filter). We note that the simpler version of the question of providing such a deterministic interpretation for the standard Kalman filter has not been addressed in the literature so far.

Proofs are contained in a full version of this paper (see Gakis and Smith (2022)) and are omitted from this extended abstract.

2. ESTIMATION PROBLEM

Consider the linear, finite-dimensional, continuous time system with the state space description:

$$\dot{x} = Ax + Bw, \tag{1}$$

$$z = Cx + Dw \tag{2}$$

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$ (full column rank) are fixed known matrices [1] and $w \in \mathcal{L}^m_{2,e}$, $x \in \mathcal{L}^n_{2,e}$ and $z \in \mathcal{L}^p_{2,e}$ are input, state and output related through this linear system. We consider the

⋆ The first author would like to acknowledge McLaren Automotive ltd for a CASE studentship awarded for doctorate research.

[1] The assumption that the system matrices are constant is for notational convenience. We note that all results are valid if the system matrices A, B, C, D are time-varying, with no change required in the proofs.

problem to estimate w and x from the measurement of the signal z, which is the same as estimating w and $x(0)$ since x is generated by (1). We assume that the state x and driving input w are not measured directly, other than (indirectly) through the measurement of z (i.e. all measurements of the system are made through the output vector z). Each element of z may correspond to an individual sensor or multiple entries of z may be generated by a single device. We will denote by $\tilde{z}$ the actual measured output signal in an experiment (see Fig. 1). We introduce the performance index:

$$C_T(w, x(0)) = \int_0^T \|\tilde{z}(t) - z(t)\|^2_{R^{-1}} dt + \|x(0) - \gamma\|^2_{\Gamma^{-1}} \tag{3}$$

where $0 < R \in \mathbb{R}^{p\times p}$, $0 < \Gamma \in \mathbb{R}^{n\times n}$, $\gamma \in \mathbb{R}^n$, $0 < T \in \mathbb{R}$ are specified. The second term on the RHS of (3) plays the role of a penalty on the deviation of the initial state from an assumed value γ. The problem we wish to solve is:

$$\inf_{w,x(0)} C_T(w, x(0)) \tag{4}$$

subject to (1) and (2). In particular we wish to compute the optimal w and $x(0)$ which we will denote by $\hat{w}$ and $\hat{x}(0)$. Using a "completion of squares" construction for the performance index (3) we can show that (4) has a unique solution (Theorem 1). First consider the system:

$$\dot{x}_1 = (A_1 - K_1C_1)x_1 + (B_1 + K_1)\tilde{z}, \tag{5}$$
$$\dot{P}_1 = A_1P_1 + P_1A_1^T - K_1RK_1^T + B_1RB_1^T, \tag{6}$$
$$K_1 = P_1C_1^TR^{-1}, \tag{7}$$
$$w_1 = D^\dagger(\tilde{z} - z_1), \tag{8}$$
$$z_1 = Cx_1 \tag{9}$$

with the initial conditions $P_1(0) = \Gamma$ and $x_1(0) = \gamma$, where we have defined:

$$A_1 = A - B_1C, \tag{10}$$
$$B_1 = BD^\dagger, \tag{11}$$
$$C_1 = (\mathrm{I} - \Pi)C, \tag{12}$$
$$\Pi = DD^\dagger, \tag{13}$$
$$D^\dagger = (D^TR^{-1}D)^{-1}D^TR^{-1}. \tag{14}$$

Then the RDE (6) has a unique positive definite solution $P_1(t) > 0$ for all $t \in [0, T]$.

Theorem 1. The optimisation problem in (4) has a unique solution $\hat{w}$, $\hat{x}(0)$ which is obtained as follows: first integrate (5)–(7) forwards in time in the interval $0 \le t \le T$ with initial conditions $P_1(0) = \Gamma$ and $x_1(0) = \gamma$; then integrate:

$$\dot{\hat{x}} = A_2\hat{x} + B_2\tilde{z}_2 \tag{15}$$

backwards in time with terminal condition:

$$\hat{x}(T) = x_1(T) \tag{16}$$

to compute $\hat{x}(0)$ (and indeed $\hat{x}$); and lastly set:

$$\hat{w} = D^\dagger(\tilde{z}_2 - C_2\hat{x}) \tag{17}$$

where:

$$A_2 = A - B_2C_2, \tag{18}$$
$$B_2 = B_1, \tag{19}$$
$$C_2 = C - RB_1^TP_1^{-1}, \tag{20}$$
$$\tilde{z}_2 = \tilde{z} - RB_1^TP_1^{-1}x_1. \tag{21}$$

Furthermore, the minimum of the performance index (3) is:

$$\inf_{w,x(0)} C_T(w, x(0)) = \int_0^T \|(\mathrm{I} - \Pi)(\tilde{z}(t) - z_1(t))\|^2_{R^{-1}} dt$$
$$= \int_0^T \|\tilde{z}(t) - \hat{z}(t)\|^2_{R^{-1}} dt + \|\hat{x}(0) - \gamma\|^2_{\Gamma^{-1}} \tag{22}$$

where we have denoted the optimal output by:

$$\hat{z} = C\hat{x} + D\hat{w}. \tag{23}$$

3. TRACKING PROBLEM

In this section we will consider a related tracking problem. Assume the state q and input u satisfy:

$$\dot{q} = Fq + Gu, \tag{24}$$
$$y = Hq + Ju \tag{25}$$

where $F \in \mathbb{R}^{n\times n}, G \in \mathbb{R}^{n\times m}, H \in \mathbb{R}^{p\times n}$ and $J \in \mathbb{R}^{p\times m}$ (full column rank) are fixed known matrices and $u \in \mathcal{L}^m_{2,e}$, $q \in \mathcal{L}^n_{2,e}$ and $y \in \mathcal{L}^p_{2,e}$ are input, state and output related through this linear system. We wish to find an input u such that the output y best tracks a desired signal $\tilde{y} \in \mathcal{L}^p_{2,e}$ over a finite horizon T together with a penalty on the deviation of the terminal state from a desired state ξ for a given but arbitrary initial state η. More precisely, we introduce the performance index:

$$W_T(u) = \int_0^T \|\tilde{y}(t) - y(t)\|^2_{R^{-1}} dt + \|q(T) - \xi\|^2_{\Xi^{-1}} \tag{26}$$

where $\xi \in \mathbb{R}^n$, $0 < \Xi \in \mathbb{R}^{n\times n}$ and propose the problem:

$$\inf_u W_T(u) \tag{27}$$

subject to (24), (25) and $q(0) = \eta$. We denote the optimal solution to (27) by $\hat{u}$. Similarly to Section 2 a completion of squares construction is used to solve (27) in Theorem 2. We first consider the system:

$$\dot{q}_1 = (F_1 + M_1H_1)q_1 + (G_1 - M_1)\tilde{y}, \tag{28}$$
$$\dot{S}_1 = F_1S_1 + S_1F_1^T + M_1RM_1^T - G_1RG_1^T, \tag{29}$$
$$M_1 = S_1H_1^TR^{-1}, \tag{30}$$
$$u_1 = J^\dagger(\tilde{y} - y_1), \tag{31}$$
$$y_1 = Hq_1 \tag{32}$$

with the terminal conditions $S_1(T) = \Xi$ and $q_1(T) = \xi$, where we have defined the matrices:

$$F_1 = F - G_1H, \tag{33}$$
$$G_1 = GJ^\dagger, \tag{34}$$
$$H_1 = (\mathrm{I} - \Lambda)H, \tag{35}$$
$$\Lambda = JJ^\dagger, \tag{36}$$
$$J^\dagger = (J^TR^{-1}J)^{-1}J^TR^{-1}. \tag{37}$$

Then the RDE (29) has a unique positive definite solution $S_1(t) > 0$ for all $t \in [0, T]$.

Theorem 2. The optimisation problem in (27) has a unique solution $\hat{u}$ which is obtained as follows: first integrate (28)–(30) backwards in time with terminal conditions $S_1(T) = \Xi$ and $q_1(T) = \xi$; then integrate:

$$\dot{\hat{q}} = F\hat{q} + G\hat{u} \tag{38}$$

forwards in time with $\hat{u}$ given by the feedback law:

$$\hat{u} = J^\dagger(\tilde{y}_2 - H_2\hat{q}) \tag{39}$$

and initial condition $\hat{q}(0) = \eta$, where we have defined:

$$H_2 = H + RG_1^TS_1^{-1}, \tag{40}$$
$$\tilde{y}_2 = \tilde{y} + RG_1^TS_1^{-1}q_1. \tag{41}$$

Furthermore, the minimum of the performance index (26) is:

$$\inf_u W_T(u) = \int_0^T \|(\mathrm{I}-\Lambda)(\tilde{y}(t)-y_1(t))\|^2_{R^{-1}}dt + \|\eta - q_1(0)\|^2_{S_1(0)^{-1}}. \quad (42)$$

4. CONSTRAINED ESTIMATION PROBLEM

We now turn our attention to the constrained optimisation problem:

$$\inf_{w,x(0)} C_T(w,x(0)) \text{ subject to } x(\tau)=\zeta \quad (43)$$

for $\zeta \in \mathbb{R}^n$ and $0 \le \tau \le T$ where (1) and (2) hold. Here $C_T(w,x(0))$ is defined as in (3) with the same meaning for $\tilde{z}$, γ and Γ. Again we wish to compute the optimal w and $x(0)$ which we will denote by $\hat{w}$ and $\hat{x}(0)$. A solution of this optimisation problem will show how the optimal cost increases compared to the unconstrained value when we demand that the state passes through a prescribed point at a given time. This will give an indication in a least squares sense of how "likely" it is that the state passes through the optimum point for the unconstrained problem. For example, if there is a sharp rise in the cost when the state is required to pass through a different point, then we may have more confidence in the value of the unconstrained optimum state at that time. We first consider the system:

$$\dot{x}_2 = A_2x_2 + B_2\tilde{z}_2, \quad (44)$$

$$\dot{P}_2 = A_2P_2 + P_2A_2^T - B_2RB_2^T, \quad (45)$$

$$w_2 = D^\dagger(\tilde{z}_2 - C_2x_2). \quad (46)$$

Theorem 3. The optimisation problem (43) has a unique solution $\hat{w}$, $\hat{x}(0)$ which is obtained as follows: first integrate (5)–(7) forwards in time in the interval $0 \le t \le T$ with initial conditions $P_1(0)=\Gamma$ and $x_1(0)=\gamma$ which gives x_1 and P_1 in the interval $0 \le t \le T$; then integrate (44)–(46) backwards in time in the interval $\tau \le t \le T$ with terminal conditions $P_2(T)=P_1(T)$ and $x_2(T)=x_1(T)$ which gives x_2 and P_2 in the interval $\tau \le t \le T$; then integrate:

$$\dot{\hat{x}} = A\hat{x} + B\hat{w} \quad (47)$$

backwards in time in the interval $0 \le t \le \tau$ with the feedback law:

$$\hat{w} = D^\dagger(\tilde{z}_2 - C_2\hat{x}); \quad (48)$$

and the terminal condition $x(\tau)=\zeta$ to find $\hat{x}$, $\hat{w}$ in the interval $0 \le t \le \tau$; then integrate (47) forwards in time in the interval $\tau \le t \le T$ with the feedback law:

$$\hat{w} = D^\dagger(\tilde{z}_3 - C_3\hat{x}) \quad (49)$$

and the initial condition $x(\tau)=\zeta$ to find $\hat{x}$, $\hat{w}$ in the interval $\tau \le t \le T$. The minimum of the performance index is:

$$\int_0^T \|(\mathrm{I}-\Pi)(\tilde{z}(t)-z_1(t))\|^2_{R^{-1}}dt + \|\zeta - x_2(\tau)\|^2_{P_2^{-1}(\tau)}. \quad (50)$$

The solution to the constrained optimisation problem (43) given in Theorem 3 introduced the vector and matrix variables x_2, w_2 and P_2. We recall that x_2 and w_2 coincide with the state and input trajectories on the interval $[0,T]$ that minimise the performance criterion (3) as shown in Theorem 1. We may now provide an interpretation of the matrix variable P_2. In Theorem 3 it is shown that the unique minimum of (43) takes the same form as the first expression on the right hand side of (22) but with an additional quadratic term which is zero when $\zeta = x_2(\tau)$, in which case we recover the results of Theorem 1. Consider now the eigenvector-eigenvalue decomposition of $P_2(\tau)$. Components of $\zeta - x_2(\tau)$ in those eigenvector directions of $P_2(\tau)$ which have small eigenvalues (i.e. large eigenvalues of $P_2^{-1}(\tau)$) give a large contribution to the second term in (50). Hence the measurements provide high confidence that the state $x(\tau)$ should be close to $x_2(\tau)$ in those directions. Fig. 2 illustrates the interpretation of $P_2(\tau)$ in the 2-D case. The figure shows an ellipse with centre $x_2(\tau)$ whose axes are aligned with the eigenvectors of $P_2(\tau)$ and lengths given by the corresponding eigenvalue square roots. All points on the ellipse increase the minimum performance index (50) by 1.

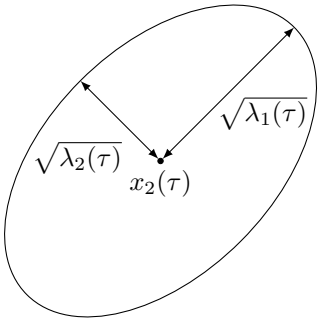


Fig. 2. An ellipse with semi-axes of length given by the eigenvalue square roots of $P_2(\tau)$, $\sqrt{\lambda_1(\tau)}$ and $\sqrt{\lambda_2(\tau)}$, and aligned with the corresponding eigenvectors.

5. SPECIAL CASES

5.1 Standard Kalman filter

We now show how the continuous time Kalman filter can be derived as a special case of the filter in Theorems 1. We therefore consider a system described by:

$$\dot{x} = Ax + Bw \quad (51)$$

$$z = Cx \quad (52)$$

with noisy measurement $\tilde{z}$ of z. Note that we assume as standard that sensor measurements of the state are not directly affected by the process noise w. In the standard Kalman filter the process noise w can be interpreted as a small magnitude disturbance to the system. To translate into our framework we need to incorporate a weighted 2-norm constraint on w in the performance index (3). In particular, we consider the following optimisation problem:

$$\inf_{w,x(0)} \left(\int_0^T \|\tilde{z}(t)-z(t)\|^2_{R^{-1}}dt + \int_0^T \|w\|^2_{Q^{-1}}dt + \|x(0)-\gamma\|^2_{\Gamma^{-1}} \right). \quad (53)$$

To translate this into the framework of this paper we introduce a virtual measurement of w which is equal to zero. More precisely, we consider an augmented system with (real and virtual) outputs given by:

$$z_a = \begin{bmatrix} C \\ 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \mathrm{I} \end{bmatrix} w \quad (54)$$

and for which we have the measurement:

$$\tilde{z}_a = \begin{bmatrix} \tilde{z} \\ 0 \end{bmatrix}. \tag{55}$$

We define an augmented block diagonal weighting matrix R_a given by:

$$R_a = \begin{bmatrix} R & 0 \\ 0 & Q \end{bmatrix}. \tag{56}$$

The following result is obtained by applying Theorem 1 to this augmented system.

Theorem 4. Consider the system:

$$\dot{x}_1 = Ax_1 + K(\tilde{z} - Cx_1), \tag{57}$$

$$\dot{P}_1 = AP_1 + P_1A^T - KRK^T + BQB^T, \tag{58}$$

$$K = P_1C^TR^{-1} \tag{59}$$

for $P_1(0) = \Gamma$ and $x_1(0) = \gamma$. The optimisation problem (53) where z is defined by (51)–(52) has a unique solution $\hat{w}$, $\hat{x}(0)$ where $\hat{w}$ is defined by the feedback law:

$$\hat{w} = QB^TP_1^{-1}(\hat{x} - x_1) \tag{60}$$

and $\hat{x}(t)$ is obtained by solving $\dot{\hat{x}} = A\hat{x} + B\hat{w}$ backwards on the interval $[0, T]$ with terminal condition $\hat{x}(T) = x_1(T)$. Furthermore, the optimal cost (53) is given by:

$$\int_0^T \|\tilde{z} - Cx_1\|^2_{R^{-1}} dt. \tag{61}$$

The filter (57)–(59) is an end-of-interval estimator in the sense that $\hat{x}(T) = x_1(T)$, $\hat{w}(T) = w_1(T)$ and takes the form of the standard Kalman filter with gain K. The above result reduces to that given in Willems (2004) with $R = \mathrm{I}$ and $Q = \mathrm{I}$.

It is interesting to note that by substituting for $\hat{w}$ from (60) we obtain an equation for the optimal state estimate:

$$\dot{\hat{x}} = A\hat{x} + BQB^TP_1^{-1}(\hat{x} - x_1) \tag{62}$$

where $\hat{x}(T) = x_1(T)$ that coincides with the standard form for the smoothed estimate in Kalman filtering (see Kailath and Frost (1968) eqn. 34(a)). Similarly by specialising (45) to the present case we have the equation:

$$\dot{P}_2 = (A + BQB^TP_1^{-1})P_2 + P_2(A + BQB^TP_1^{-1})^T - BQB^T \tag{63}$$

where $P_2(T) = P_1(T)$, which is the corresponding form for the smoothed covariance (see Kailath and Frost (1968) eqn. 34(b)).

5.2 Standard LQR on a finite time horizon

We now show how the solution of the standard linear quadratic regulator (LQR) on a finite time horizon can be derived as a special case of the tracking problem in Theorem 2. In the standard LQR we wish to find a low energy input u that brings the state q to the origin. More precisely, we consider the optimisation problem:

$$\inf_u \left(\int_0^T \|q(t)\|^2_{R_q^{-1}} + \|u(t)\|^2_{R_u^{-1}} dt + \|q(T)\|^2_{\Xi^{-1}} \right) \tag{64}$$

where the state q and input u satisfy:

$$\dot{q} = Fq + Gu \tag{65}$$

and the initial state $q(0) = \eta$ is known. To put this into the form required to apply Theorem 2 we choose:

$$y = \begin{bmatrix} \mathrm{I} \\ 0 \end{bmatrix} q + \begin{bmatrix} 0 \\ \mathrm{I} \end{bmatrix} u, \tag{66}$$

$\tilde{y} = 0$, $\xi = 0$ and:

$$R = \begin{bmatrix} R_q & 0 \\ 0 & R_u \end{bmatrix}. \tag{67}$$

Theorem 5. Consider the RDE:

$$\dot{S}_1 = FS_1 + S_1F^T + S_1R_q^{-1}S_1^T - GR_uG^T \tag{68}$$

with the terminal condition $S_1(T) = \Xi$. The optimisation problem (64) has a unique solution $\hat{u}$ which is defined by the feedback law:

$$\hat{u} = -R_uG^TS_1^{-1}\hat{q} \tag{69}$$

and $\hat{q}(t)$ is obtained by solving $\dot{\hat{q}} = F\hat{q} + G\hat{u}$ forwards in the interval $[0, T]$ with the initial condition $\hat{q}(0) = \eta$. Furthermore, the optimal cost (64) is:

$$\|\eta\|^2_{S_1(0)^{-1}}. \tag{70}$$

This is recognised as the classical LQR result on a finite time horizon.

6. CONCLUSIONS

The paper has proposed a framework for estimation in which the output of the dynamical system comprises *all* variables that are measured, and the variables to be estimated comprise, equally, system states and exogenous inputs. This framework is quite general in that if an exogenous input is measured then we may include a direct feedthrough component in the output vector to reflect this. This estimation problem was solved for linear systems with a full column rank feedthrough matrix. The unique optimum solution on a finite horizon takes a two-stage form in which the first stage provides an end-of-interval estimator which can be solved in real time as the horizon length increases. The full rank assumption is general enough to include the Kalman filter and, for the dual tracking problem, the linear quadratic regulator as special cases. Generalising the result to the case where this condition, or similar full rank assumptions on the Markov parameters, does not hold remains an open problem.

A contribution of this paper has been to provide an interpretation of the solution of the Riccati differential equation which is analogous to the meaning of the covariance matrix in stochastic filtering. The solution of a matrix Lyapunov differential equation $P_2(t)$ is shown to have an analogous interpretation to the smoothed covariance in the stochastic case. This has been achieved by considering the least-squares estimation problem with an additional constraint that the state passes through a prescribed point at a given time in the fixed horizon. To solve this problem a tracking problem was also considered which is dual to our estimation problem.

REFERENCES

Gakis, G. and Smith, M.C. (2022). A deterministic least squares approach for simultaneous input and state estimation. *Submitted to IEEE Transactions on Automatic Control.*

Kailath, T. and Frost, P. (1968). An innovations approach to least-squares estimation–part II: Linear smoothing in additive white noise. *IEEE Transactions on Automatic Control*, 13(6), 655–660.

Kalman, R.E. and Bucy, R.S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering, Transactions of the ASME*, 83(3), 95–108.

Willems, J.C. (2004). Deterministic least squares filtering. *Journal of econometrics*, 118(1-2), 341–373.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Security considerations about a McEliece cryptosystem with a convolutional encoder ⋆

P. Almeida * **M. Beltra** ** **D. Napp** *** **C. Sebastião** ****

* *Dep. de Matemática, Universidade de Aveiro 3810-193 Aveiro, Portugal (e-mail: palmeida@ua.pt).*
** *Dep. de Matemàtiques, Universitat d'Alacant 99 e-03080. Alicante. Spain (e-mail: miguelbeltravidal@gmail.com).*
*** *Dep. de Matemàtiques, Universitat d'Alacant 99 e-03080. Alicante. Spain (e-mail: diego.napp@ua.es).*
**** *Dep. de Matemática, Universidade de Aveiro 3810-193 Aveiro, Portugal (e-mail: claudia.sebastiao@ua.pt).*

Abstract: We present a new variant of the McEliece cryptosystem that possesses several interesting properties, including a reduction of the public key for a given security level. In contrast to the classical McEliece cryptosystems, where block codes are used, we propose the use of a convolutional encoder to be part of the public key. The secret key is constituted by a Generalized Reed-Solomon code and two Laurent polynomial matrices that contain large parts that are generated completely at random. In this setting the message is a sequence of messages instead of a single block message and the errors are added randomly throughout the sequence. We analyse its security against ISD attacks in the first instants and when the whole message is transmitted, as well as against structural attacks.

Keywords: Convolutional codes, Generalized Reed-Solomon codes, McEliece Cryptosystem, Information Set Decoding.
AMS Subject Classification: 94A60, 94B10, 11T71.

1. INTRODUCTION

Code-based cryptosystems are considered promising alternatives for Public Key Cryptography (PKC) since their security relies on well-known NP-hard problems and allow fast encryption and decryption procedures. Moreover, they are immune to attacks that use Shor's algorithm and therefore are candidates for post-quantum cryptography. However, one of the main disadvantages of code-based schemes is the large keys whose size is inherently determined by the underlying Goppa block code used in the original cryptosystem. For this reason, there have been several attempts to substitute Goppa codes by other classes of block codes, e.g. Generalized Reed-Solomon (GRS) codes among many others. GRS codes are Maximum Distance Separable (MDS) codes which, in the McEliece scheme, translates into smaller key sizes. Hence, a major improvement would be achieved if these codes could be securely used in these cryptosystems. Unfortunately, due to their strong algebraic structure, they are vulnerable to many structural attacks. One recent interesting idea to remove this algebraic structure was to replace the permutation used in the original McEliece cryptosystem with a more general transformation (see Baldi et al. (2016)). However, this variant has also been fully broken using an attack based on the Schur square code distinguisher that permit to distinguish GRS codes from random ones in Couvreur et al. (2015) and Couvreur and Lequesne (2020).

We continue this line of research and explore a new variant that allows the use of GRS codes using a convolutional mask. In our scheme, the plaintext is not a block vector but a stream of smaller vectors sent in a sequential fashion. The public key is given by the polynomial convolutional encoder $G'(D) = S(D)GP(D^{-1}, D)$ where G is the generator matrix of a GRS, $S(D)$ a polynomial matrix and $P(D^{-1}, D)$ an invertible Laurent polynomial matrix. Hence, the proposed class of convolutional codes uses GRS codes and adds a convolutional layer to it in order to thwart the key recovery attack against GRS codes and at the same time admits a simple iterative algebraic decoding algorithm. This idea is different from the above ideas and therefore the cryptanalysis has to be adapted to this case.

The matrices $S(D)$ and $P(D^{-1}, D)$ are selected to protect against both ISD and structural attacks. A crucial fact to ensure security against ISD attacks to the first blocks and bootstrap from there, is that the truncated sliding matrices of the convolutional encoder are not full row rank and have distance equal to one, so recovering the initial blocks is not

⋆ This work was supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT), references UIDB/04106/2020. The second and third authors were supported by Spanish grants PID2019-108668GB-I00 of the ministerio de Ciencia e Innovación of the Gobierno de España and VIGROB-287 of the Universitat d'Alacant.

possible. As for structural attacks, we build our matrix $P(D^{-1}, D)$ to provide weight-2 masking at every instant. As pointed out in Bolkema et al. (2017); Khathuria et al. (2018), (see also Khathuria et al. (2021)), the weight-two masking appears to be enough to remove any identifiable algebraic structure from the public code and therefore structural attacks, and in particular any distinguisher attack based on the Schur product, seems to fail as well. We note that our construction uses matrices with large parts generated completely at random and can be easily constructed. We present several examples for comparison with previous variants of the McEliece cryptosystem to illustrate the key size reduction of the public key achieved by this novel scheme.

2. A MCELIECE CRYPTOSYSTEM WITH A CONVOLUTIONAL ENCODER

For the purposes of this work we need to construct convolutional codes with certain properties that will allow the cryptosystem to work efficiently and resist cryptanalysis. The ingredients for building up our cryptosystem are the following: let F be a finite field with q elements, $G \in F^{k\times n}$ be an encoder of an (n, k) block code that admits an efficient decoding algorithm and can correct up to t errors,

$$S(D) = S_1 D + S_2 D^2 \in F^{k\times k}[D] \tag{1}$$

where $S_1 \in F^{k\times k}$ is an invertible constant matrix and $S_2 \in F^{k\times k}$ is generated at random and

$$T(D^{-1}, D) = T_{-1}D^{-1} + T_0 + T_1 D \in F^{n\times n}[D^{-1}, D] \tag{2}$$

is an invertible (in $F(D)$) Laurent polynomial matrix with:

(i) The determinant satisfies $|T(D^{-1}, D)| \in F \setminus \{0\}$.
(ii) Nonzero rows of T_j have 2 nonzero elements, for $j \in \{-1, 0, 1\}$.
(iii) Indices of nonzero columns of T_{-1}, T_0 and T_1 form a partition of n.
(iv) If we denote

$$T^{-1}(D^{-1}, D) = P(D^{-1}, D) = P_{-1}D^{-1} + P_0 + P_1 D,$$

then, the nonzero columns of each matrix P_{-1}, P_0 and P_1, have at least 2 nonzero entries.

With them, we construct the convolutional encoder

$$\begin{aligned} G'(D) &= S(D)\ G\ P(D^{-1}, D) \\ &= G'_0 + G'_1 D + G'_2 D^2 + G'_3 D^3. \end{aligned} \tag{3}$$

Let $\mathrm{wt}(\mathbf{v})$ be the Hamming weight of a vector $\mathbf{v} \in F^n$, i.e. the number of the nonzero components of $\mathbf{v}$. This definition can be extended to polynomial vectors $\mathbf{v}(D^{-1}, D) = \sum_{i\in Z} \mathbf{v}_i D^i$ in a natural way as $\mathrm{wt}(\mathbf{v}(D^{-1}, D)) = \sum_{i\in Z} \mathrm{wt}(\mathbf{v}_i)$.

For the sake of simplicity we shall consider information vectors that start at time instant zero and have finite support, i.e., $\mathbf{u}(D) \in F^k[D]$ is polynomial. We will also consider the error vectors $\mathbf{e}(D) \in F^n[D]$ to be polynomial.

Lemma 1. Let $T(D^{-1}, D)$ be as described above and $\mathbf{e}(D) = \sum_{i\geq 0} \mathbf{e}_i D^i \in F^n[D]$ a random vector satisfying

$$\mathrm{wt}([\mathbf{e}_i\ \mathbf{e}_{i+1}\ \mathbf{e}_{i+2}]) \leq \frac{t}{2} \tag{4}$$

for all $i \geq 0$. Then all the coefficients of $\mathbf{e}(D)T(D^{-1}, D)$ have weight less than or equal to t.

Proof. The coefficient of degree ℓ of $\mathbf{e}(D)T(D^{-1}, D)$ is

$$T_{-1}\mathbf{e}_{\ell+1} + T_0\mathbf{e}_\ell + T_1\mathbf{e}_{\ell-1.},$$

where $\mathbf{e}_i = 0$ for $i < 0$. Since each row of T_j has at most 2 nonzero elements, then $\mathrm{wt}(\mathbf{e}_i T_j) \leq 2\mathrm{wt}(\mathbf{e}_i)$ for all $i \geq 0$ and $j \in \{-1, 0, 1\}$, and the result follows.

Condition (4) describes the maximum number of errors allowed within a time interval and is similar to the sliding window condition introduced in Badr et al. (2017) to describe the possible error patterns that can occur in a given channel.

Theorem 2. Let $G'(D)$ be the encoder as described in (3), t the correcting error capability of G,

$$\mathbf{u}(D) = \mathbf{u}_0 + \mathbf{u}_1 D + \cdots + \mathbf{u}_s D^s$$

the information sequence and

$$\mathbf{e}(D) = \mathbf{e}_0 + \mathbf{e}_1 D + \cdots + \mathbf{e}_{s+3} D^{s+3}$$

an error vector satisfying (4). Then, the received data $\mathbf{y}(D) = \mathbf{u}(D)G'(D) + \mathbf{e}(D) \in F^n[D]$ can be decoded.

Proof. Multiplying $\mathbf{y}(D)$ by $T(D^{-1}, D)$ from the right yields the polynomial equation

$$\mathbf{y}(D)T(D^{-1}, D) = \mathbf{u}(D)S(D)G + \mathbf{e}(D)T(D^{-1}, D).$$

Hence, for some coefficients $\widehat{\mathbf{u}}_i \in F^k$, we can write

$$\mathbf{u}(D)S(D) = \sum_{i=1}^{s+2} \widehat{\mathbf{u}}_i D^i \tag{5}$$

and for some coefficients $\widehat{\mathbf{e}}_i \in F^n$, we can write

$$\mathbf{e}(D)T(D^{-1}, D) = \sum_{i=-1}^{s+4} \widehat{\mathbf{e}}_i D^i,$$

and therefore each coefficient of $\mathbf{y}(D)T(D^{-1}, D)$ is of the form $\widehat{\mathbf{u}}_i G + \widehat{\mathbf{e}}_i$. By Lemma 1 it follows that $\mathrm{wt}(\widehat{\mathbf{e}}_i) \leq t$, for $-1 \leq i \leq s + 4$ and, therefore, each $\widehat{\mathbf{u}}_i$ can be recovered. Further, from (5) we have that

$$[\,\widehat{\mathbf{u}}_1\ \widehat{\mathbf{u}}_2\ \cdots\ \widehat{\mathbf{u}}_{s+2}\,] = [\,\mathbf{u}_0\ \mathbf{u}_1\ \cdots\ \mathbf{u}_s\,] \begin{bmatrix} S_1 & S_2 & & & \\ & S_1 & S_2 & & \\ & & \ddots & \ddots & \\ & & & S_1 & S_2 \end{bmatrix},$$

so, one can recover each $\mathbf{u}_i$ sequentially as

$$\begin{aligned} \mathbf{u}_0 &= \widehat{\mathbf{u}}_1 S_1^{-1}, \\ \mathbf{u}_1 &= (\widehat{\mathbf{u}}_2 - \mathbf{u}_0 S_2)\, S_1^{-1}, \\ \mathbf{u}_2 &= (\widehat{\mathbf{u}}_3 - \mathbf{u}_1 S_2)\, S_1^{-1}, \\ &\vdots \\ \mathbf{u}_i &= (\widehat{\mathbf{u}}_{i+1} - \mathbf{u}_{i-1} S_2)\, S_1^{-1}. \end{aligned}$$

The proposed scheme works as follows:

Secret key: Generate a triple $\{S(D), G, T(D^{-1}, D)\}$, compute $G'(D) = G'_0 + G'_1 D + G'_2 D^2 + G'_3 D^3$ and define the following sets

$$N_j = \{\text{indices of the null columns of } G'_j\}, \ \ j \in \{0, 1, 2, 3\}.$$

$$J_0 = \{1, \ldots, n\} \setminus N_0, \ J_1 = N_0 \setminus N_1, \ J_2 = (N_0 \cap N_1) \setminus N_2.$$

Let $\widetilde{G}_j$ be the submatrix of G'_j whose columns are indexed by J_j, and define $\widetilde{G} = [\widetilde{G}'_0 | \widetilde{G}'_1 | \widetilde{G}'_2]$. Let $E(G'_0)$ and $E(\widetilde{G})$ be the reduced row echelon form of G'_0 and $\widetilde{G}$ and check if the following conditions are satisfied:

(a) $E(G'_0)$ has rows of weight 1.
(b) $E(\widetilde{G})$ has rows of weight 1.

If this is the case, then $\{S(D), G, T(D^{-1}, D)\}$ is correctly generated. If not, generate another instance of $S(D)$, G and $T(D^{-1}, D)$.

Remark 3. After randomly generating a sample of 500 matrices $T(D^{-1}, D)$ of the form described in Section 2.1, we found that about a 4% of them fulfill **(a)** and **(b)**.

Public key: $\{G'(D) = S(D)GT^{-1}(D^{-1}, D), \ t/2\}$.

Encryption: Let $\pi_I(\mathbf{v})$ be the projection of $\mathbf{v}$ into the coordinates in I and let us define the following sets:

$$L = \{\text{column indices of } [G'_0 | G'_1 | G'_2] \text{ where the rows of weight 1 of } E(\widetilde{G}) \text{ have the nonzero elements}\},$$

$$L' = \{\text{column indices of the nonzero elements of the rows of weight 1 of } E(G'_0)\}.$$

To encrypt a message

$$\mathbf{u}(D) = \mathbf{u}_0 + \mathbf{u}_1 D + \mathbf{u}_2 D^2 + \cdots + \mathbf{u}_s D^s \in F^k[D],$$

Alice selects an error vector

$$\mathbf{e}(D) = \mathbf{e}_0 + \mathbf{e}_1 D + \cdots + \mathbf{e}_{s+3} D^{s+3} \in F^n[D],$$

satisfying (4) and the following additional conditions:

$$\mathrm{wt}(\pi_{L'}(\mathbf{e}_0)) = \mathrm{wt}(\mathbf{e}_0), \tag{6}$$

$$\mathrm{wt}(\pi_{J_2}(\mathbf{e}_1)) = 0, \tag{7}$$

$$\mathrm{wt}(\pi_L([\mathbf{e}_0 \ \mathbf{e}_1 \ \mathbf{e}_2])) = \mathrm{wt}([\mathbf{e}_0 \ \mathbf{e}_1 \ \mathbf{e}_2]), \tag{8}$$

$$\mathrm{wt}(\pi_L([\mathbf{e}_i \ \mathbf{e}_{i+1} \ \mathbf{e}_{i+2}]) \geq 1, \quad 0 \leq i \leq s+1 \tag{9}$$

$$\mathrm{wt}(\pi_{L'}(\mathbf{e}_i)) \geq 1, \quad 0 \leq i \leq s \tag{10}$$

and encrypts the message as

$$\mathbf{y}(D) = \mathbf{u}(D)G'(D) + \mathbf{e}(D). \tag{11}$$

Decryption: Bob multiplies (11) from the right by the matrix $T(D^{-1}, D)$ to obtain

$$\mathbf{u}(D)S(D)G + \mathbf{e}(D)T(D^{-1}, D), \tag{12}$$

he decodes using G and recovers the message $\mathbf{u}(D)$ from $\mathbf{u}(D)S(D)$, as explained in the proof of Theorem 2.

2.1 Constructing $T(D^{-1}, D)$

In this subsection we present a large family of matrices $T(D^{-1}, D)$ satisfying properties **(i)**, **(ii)**, **(iii)** and **(iv)**. To this end, we first recall a technical lemma about the determinant and inverse of a block matrix, first obtained by I. Schur (Schur (1917)).

Lemma 4. (Cottle, 1974, Formulas (2) and (4)) Let T be a block matrix of the form

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where A and D are nonsingular. Then

a) $|T| = |A| \ |D - CA^{-1}B|$.
b) If T is invertible, the inverse of T is

$$P = -\begin{bmatrix} -(A - BD^{-1}C)^{-1} & A^{-1}B(D - CA^{-1}B)^{-1} \\ D^{-1}C(A - BD^{-1}C)^{-1} & -(D - CA^{-1}B)^{-1} \end{bmatrix}.$$

Starting from a 2×2 block matrix facilitates the construction of a matrix satisfying **(ii)**. Moreover, since the above lemma tells us how the determinant and the inverse matrix are, properties **(i)** and **(iv)** can be easily achieved as well. The blocks we use to construct $T(D^{-1}, D)$ are denoted by $A(D^{-1}, D)$ and are of size $\frac{n}{2} \times \frac{n}{2}$, where n is the size of the code (we use even values for n). These blocks are randomly generated satisfying the following properties:

Properties 5. $A = A(D^{-1}, D) \in F^{\frac{n}{2} \times \frac{n}{2}}[D^{-1}, D]$ satisfies:

(1) A is an upper triangular matrix;
(2) The entries of the principal diagonal of A are of the form aD^j, with $a \in F \setminus \{0\}$, and $j \in \{-1, 0, 1\}$ in such a way that there are δ_j entries with power D^j, satisfying
$$\delta_{-1} = \delta_1; \tag{13}$$
(3) Each row of A has at most one entry of the form γD^j for each $j \in \{-1, 0, 1\}$, with $\gamma \in F \setminus \{0\}$;
(4) All nonzero entries of a column of A have the same exponent of D.

Condition (13) can be understood as the sum of the exponents of D along the diagonal is 0. Since A is upper triangular and the diagonal entries are nonzero, this implies that $|A| \in F \setminus \{0\}$, i.e., the determinant is a nonzero constant. Once we have determined how the blocks we will use are, we can construct $T(D^{-1}, D)$ satisfying properties **(i)**, **(ii)**, **(iii)** and **(iv)**, as it shown the next theorem.

Lemma 6. Suppose $|F| > 2$. Let $\Gamma \in F^{n \times n}$ be a permutation matrix, $A = A(D^{-1}, D)$ be a matrix satisfying Properties 5, take $\beta \in F \setminus \{0, 1\}$ and consider $\Delta(D^{-1}, D)$ to be the block matrix

$$\Delta(D^{-1}, D) = \begin{bmatrix} A & \beta A \\ A & A \end{bmatrix} \in F^{n \times n}[D^{-1}, D].$$

Then $T(D^{-1}, D) = \Gamma \, \Delta(D^{-1}, D)$ satisfies properties **(i)**, **(ii)**, **(iii)** and **(iv)**.

Proof. Using Lemma 4 **a)**, we have

$$|\Delta(D^{-1}, D)| = |A| \, |A - AA^{-1}(\beta A)| \ = |A|^2 |I - \beta I|.$$

Since $|A| \in F \setminus \{0\}$ and $\beta \neq 1$ then $|T(D^{-1}, D)| \in F \setminus \{0\}$, so condition **(i)** holds.

Conditions **(ii)** and **(iii)** hold due to Properties 5 and our construction of $\Delta(D^{-1}, D)$. Condition **(iv)** is obtained using Lemma 4 **b)** since we have

$$\Delta^{-1}(D^{-1}, D) = \begin{bmatrix} \frac{1}{1-\beta}A^{-1} & \frac{\beta}{\beta-1}A^{-1} \\ \frac{1}{\beta-1}A^{-1} & \frac{1}{1-\beta}A^{-1} \end{bmatrix}.$$

If we write $\Delta(D^{-1}, D) = \sum_{j=-1}^{1} \Delta_j D^j$ then, by construction, all rows of Δ_j, and hence all columns of P_j, will have an even number of nonzero entries, for $j \in \{-1, 0, 1\}$.

3. ATTACKS AGAINST THE PROPOSED CRYPTOSYSTEM

In this section, we briefly comment some of possible attacks to the proposed cryptosystem and show how the properties analyzed in previous sections provide increased key security with respect to previous variants of the McEliece cryptosystem.

An attacker could consider the full row rank matrix

$$\begin{bmatrix} G'_0 & G'_1 & G'_2 & G'_3 & & & \\ & G'_0 & G'_1 & G'_2 & G'_3 & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & G'_0 & G'_1 & G'_2 & G'_3 \end{bmatrix}$$

as the generator matrix of a block code. Note, however, that the size of this block code is $k(s+1) \times n(s+4)$ and therefore it will be very large for large values of s, even if n and k are small. Table 1 illustrates a few specific examples with the work factor (WF) needed to recover $\mathbf{u}(D)$ using ISD attacks on that matrix (see Bolkema et al. (2017) for more details on the computation of the WF).

n	k	s	WF (bin. op.)	Public Key (bits)
New Scheme				
108	60	24	$2^{129.3}$	182695
108	60	51	$2^{256.7}$	182696
180	108	14	$2^{128.5}$	624713
180	108	31	$2^{262.6}$	624714
180	108	62	$2^{512.6}$	624715
180	114	14	$2^{132.0}$	659417
180	114	30	$2^{260.6}$	659418
180	114	61	$2^{513.2}$	659419
Classical McEliece with Goppa codes				
2960	2288		2^{128}	1537536
8192	6528		2^{256}	10862592
GRS with 2-weight mask (Bolkema et al.)				
764	496		$2^{128.10}$	1329280
762	490		$2^{128.02}$	1332800
766	498		$2^{128.35}$	1334640

Table 1. Parameters, work force and public key sizes of PKC

An attacker could also try to use the convolutional structure of the scheme and sequentially decode each block. Notice that if the attacker has a method to recover $\mathbf{u}_0$ and $\mathbf{e}_0$, then by computing $D^{-1}(\mathbf{y}(D) - \mathbf{u}_0 G'(D) - \mathbf{e}_0)$ the process can be repeated to recover the rest of the $\mathbf{u}_i$, $\mathbf{e}_i$, for $i \geq 1$. For this purpose the attacker can start with the block code $\mathcal{C}_0$ associated with G'_0. However, since the distance of the code $\mathcal{C}_0$ is 1, the number codewords in a ball of radius $\leq t$ centered in $\mathbf{y}_0$ is expected to be really huge. Moreover, condition (6) implies that $\mathbf{y}_0$ is indeed a codeword of $\mathcal{C}_0$.

Another way to try to recover $\mathbf{u}_0$ is to consider all the equations involving $\mathbf{u}_0$ and none of the other $\mathbf{u}_i$, for $i \geq 1$. This system of equations is precisely given by the matrix $\widetilde{G}$ and the associated parts of the vector $[\mathbf{y}_0\ \mathbf{y}_1\ \mathbf{y}_2]$. Again, since the distance of the code generated by the matrix $\widetilde{G}$ has distance 1, it is expected to have a large list of codewords centered in that vector and no methods to decide which is the correct one.

For the structural attacks, since the weight of the nonzero columns of P_i, with $i \in \{-1, 0, 1\}$, is larger than or equal to 2 due to condition **(iv)**, the attacks presented in Couvreur et al. (2015) and Couvreur and Lequesne (2020) don't seem to succeed against the proposed scheme.

ACKNOWLEDGEMENTS

The authors would like to thank Karan Khathuria and Jean-Pierre Tillich for useful comments on this work. This work was supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e Tecnologia), references UIDB/04106/2020. The second and third authors were supported by Spanish grants PID2019-108668GB-I00 of the Ministerio de Ciencia e Innovación of the Gobierno de España and VIGROB-287 of the Universitat d'Alacant.

REFERENCES

Badr, A., Khisti, A., Tan, W.T., and Apostolopoulos, J. (2017). Layered constructions for low-delay streaming codes. 63(1), 111–141.

Baldi, M., Bianchi, M., Chiaraluce, F., Rosenthal, J., and Schipani, D. (2016). Enhanced public key security for the McEliece cryptosystem. *Journal of Cryptology*, 29(1), 1–27.

Bolkema, J., Gluesing-Luerssen, H., Kelley, C.A., Lauter, K.E., Malmskog, B., and Rosenthal, J. (2017). Variations of the McEliece cryptosystem. In J.W. E. Howe K. Lauter (ed.), *Algebraic Geometry for Coding Theory and Cryptography. Association for Women in Mathematics Series*, volume 9. Springer Cham.

Cottle, R.W. (1974). Manifestations of the Schur complement. *Linear Algebra Appl.*, 8, 189–211.

Couvreur, A. and Lequesne, M. (2020). On the security of subspace subcodes of Reed-Solomon codes for public key encryption. *arXiv: https://arxiv.org/abs/2009.05826.*

Couvreur, A., Otmani, A., Tillich, J.P., and Gauthier-Umana, V. (2015). A polynomial-time attack on the BBCRS scheme. *Conference Public Key Cryptography (PKC), 2015 https://arxiv.org/pdf/1501.03736.pdf.*

Khathuria, K., Rosenthal, J., and Weger, V. (2018). Two masking of the Reed-Solomon structure in conjugation with list decoding. In *Proceedings of 23rd International Symposium on Mathematical Theory of Networks and Systems, Hong Kong University of Science and Technology, Hong Kong*, 309–314.

Khathuria, K., Rosenthal, J., and Weger, V. (2021). Encryption scheme based on expanded Reed-Solomon codes. *Advances in Mathematics of Communications*, 15(2), 207–218.

Schur, I. (1917). Potenzreihen im Innern des Einheitskreises. *J. Reine Angew. Math.*, 147, 205–232.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Mechanical realisation of a lossless adjustable two-port transformer ⋆

Lukas Gaudiesius * **Tryphon T. Georgiou** ** **Neil E. Houghton** ***
Malcolm C. Smith ****

* *Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: lg608@cam.ac.uk).*
** *Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA 92697, U.S.A. (e-mail: tryphon@uci.edu).*
*** *Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: neh27@eng.cam.ac.uk).*
**** *Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: mcs@eng.cam.ac.uk).*

Abstract: This paper continues the work of Georgiou, Jabbari and Smith on lossless adjustable mechanical devices. Defining equations and mechanical constructions of lossless adjustable springs and inerters for translational and rotational devices will be recalled. The role played by the lossless adjustable two-port transformer will be highlighted. A mechanical design will be described for a lossless adjustable rotational two-port transformer involving a double-cone arrangement, movable carriage and a pair of counter-rotating balls.

Keywords: Passivity-based control; Mechanical networks.

1. INTRODUCTION

In Georgiou et al. (2020) the question was posed as to whether it is possible to build lossless adjustable springs and inerters. A lossless adjustable spring would have a "workless knob" and would behave like a conventional linear spring when the knob is stationary. Energy imparted through compression or extension would be available for extraction again. Adjustment of the knob would not involve any energy transfer between the environment and the contrivance. Current methods to adjust the stiffness of springs do not answer this question, since they require active actuation, dissipation, or restrictive conditions on the switching of the spring constant.

The question is motivated by the ubiquity of the adjustable damper which is extensively used in the control of mechanical systems, e.g. automotive suspensions, see Butsuen and Hedrick (1989), Savaresi et al. (2010), Brezas et al. (2015), Smith et al. (2018). The variable damper constant plays the role of a control input which may be adjusted by a control law that minimises a performance criterion. Such devices are sometimes termed "semi-active" since a (small) power source is employed to effect the adjustment. Nevertheless, the instantaneous power absorbed by the device can never be negative, and so from a terminal point of view it appears passive.

An analogous question arises for the inerter which is a two-terminal mechanical device such that the equal and opposite force at the terminals is proportional to the relative acceleration between them (see Smith (2002), Smith (2020)). The constant of proportionality is termed the inertance. The question is whether an adjustable inerter is physically realisable as a lossless device, i.e. whether an inerter can be manufactured with a "workless knob" which freely adjusts its inertance in real time.

In the robotics field "Variable Stiffness Actuators" have been considered extensively (see Vanderborght et al. (2013), Wolf et al. (2016) for recent surveys and the references therein). Each of the methods described requires some form of active force input, most commonly via electromechanical actuation. In Bobrow et al. (1995), Jabbari and Bobrow (2002) a passive "resettable" spring is proposed which requires minimal energy for switching. The opening of the valve (to reduce the stiffness) is constrained to times at which there is no stored energy in the fluid, otherwise there is energy dissipation. The possible benefits of adjustable inerters have been considered recently in Chen et al. (2014), Brzeski et al. (2015), Lazarek et al. (2018), Garrido et al. (2018) without identifying a method to make the adjustments in a lossless manner.

The present extended abstract is structured as follows. Section 2 highlights the mechanical constructions of Georgiou et al. (2020): an idealised mechanical arrangement of a lever with movable fulcrum to derive device laws for lossless adjustable springs and inerters. The definitions of the varspring and varinerter will be recalled, as well as the rotational counterparts of these concepts. Section 3 focusses specifically on the rotary adjustable transformer and details a specific embodiment of the physical realisation of the device.

2. PLANAR MECHANISM FOR LOSSLESS ADJUSTABLE DEVICES

2.1 Lossless adjustable spring

We consider a theoretical mechanism as depicted in Fig. 1 in which the x- and y-axes are fixed in the device housing. The device terminals are located at $(0,0)$ and $(x_1,0)$ according to the convention of Georgiou et al. (2020) and accordingly we

⋆ L. Gaudiesius acknowledges the support of an EPSRC doctoral studentship. Partial support was provided by the NSF (1665031, 1807664, 1839441) and AFOSR (FA9550-17-1-0435).

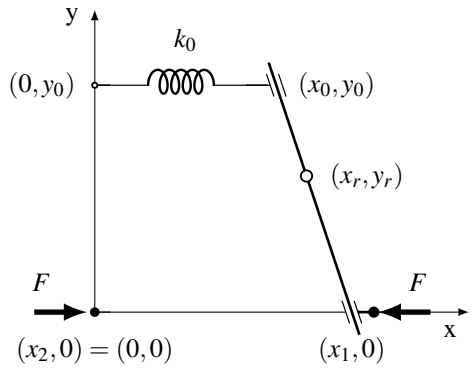


Fig. 1. Spring and lever with movable fulcrum at (x_r, y_r).

define $x = -x_1$ and $v = -\dot{x}_1$. An internal spring with stiffness k_0 is constrained to move parallel to the x-axis with fixed y-coordinate y_0 and generates a force equal to $-k_0 x_0$. An ideal massless lever has a movable fulcrum at (x_r, y_r).

If the fulcrum is movable with an imposed condition that the instantaneous power supplied at the external terminals of the device equals the rate of change of the internal energy of the spring then it can be shown as in Georgiou et al. (2020) that the fulcrum must always move parallel to the bar. Setting $k_0 = 1$ it can be shown that the device is governed by the relation:

$$\dot{x} = p\frac{d}{dt}(pF). \tag{1}$$

where

$$p = y_r/(y_0 - y_r). \tag{2}$$

We can see that $F\dot{x} = \frac{d}{dt}(\mathscr{I})$ where we may define the internal stored energy by:

$$\mathscr{I} = \frac{1}{2}p^2F^2.$$

2.2 Lossless adjustable inerter

We consider the mechanism as depicted in Fig. 2 which is similar to the device in Fig. 1 except that the spring is replaced by an inerter which generates a force equal to $-b\ddot{x}_0$.

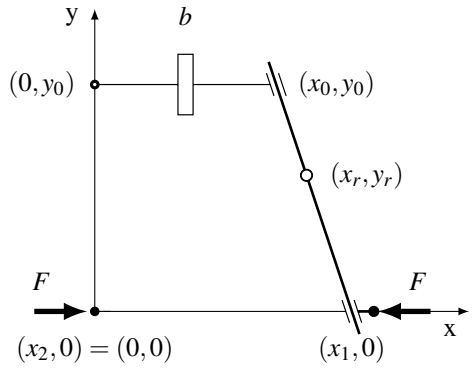


Fig. 2. Inerter and lever with movable fulcrum at (x_r, y_r).

Applying again the condition that the instantaneous power supplied at the external terminals of the device equals the rate of change of the internal energy of the inerter gives, for $b = 1$,

$$F = r\frac{d}{dt}(r\dot{x}) \tag{3}$$

where $r = p^{-1}$. It is immediate to see that $F\dot{x} = \frac{d}{dt}(\mathscr{I})$ where we may define the internal stored energy by:

$$\mathscr{I} = \frac{1}{2}r^2\dot{x}^2.$$

2.3 Physical implementation

A conceptual scheme to realise such adjustability is shown in Fig. 3. A wheel is attached to the bar at the fulcrum and is free to rotate about a vertical axis through the fulcrum and the contact point of the wheel on a supporting table. The wheel is allowed to rotate about a horizontal axis which is perpendicular to the bar to produce a rolling motion on the table which is always instantaneously parallel to the bar. The rolling of the wheel is the means of mechanism adjustment by altering the ratio r or $p = r^{-1}$ with p defined as in (2).

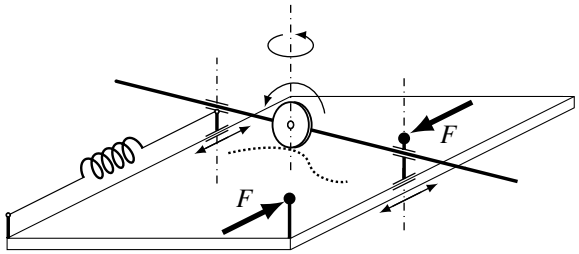


Fig. 3. Schematic of a lever mechanism with movable fulcrum to allow a physical realisation of lossless adjustable springs and inerters.

2.4 The varspring and varinerter

Based on the construction of Fig. 3, it appears justified to introduce a pair of ideal, lossless adjustable mechanical one-ports which we will name the *varspring* and *varinerter*. The ideal devices are defined by the laws:

$$v = p\frac{d}{dt}(pF) \qquad \text{(varspring)} \tag{4}$$

$$F = r\frac{d}{dt}(rv) \qquad \text{(varinerter)} \tag{5}$$

where (F, v) is the force-velocity pair of the mechanical one-port and $p(t)$, $r(t)$ are positive and freely adjustable parameters. The internal energy of the devices is given by $\frac{1}{2}p^2F^2$ and $\frac{1}{2}r^2v^2$ respectively. It is important that physical devices may be constructed which approximate the ideal behaviour, for example, having sufficiently small dissipation through friction, and as in the case of the ideal inerter in Smith (2002), sufficiently small mass, sufficient travel, have no physical attachment to a fixed point in space, and have two terminals which are freely and independently movable (Smith, 2002, Section II.C). The construction of Fig. 3 suggests that devices satisfying these conditions are physically realisable in principle. The varinerter is realised as in Fig. 3 with an inerter replacing the spring. We note that the above construction of the varspring and varinerter in Fig. 3 can be conceptualized as a lossless adjustable two-port transformer with one of the ports terminated with either a spring or an inerter.

2.5 A lossless adjustable transformer

Motivated by the method of constructing the translational varspring and varinerter in Sections 2.1 and 2.2 we consider now the possibility of an adjustable rotary transformer. We consider the construction depicted in Fig. 4 consisting of two right circular cones of equal aperture on parallel rotating shafts,

with opposite orientation, and hence a constant perpendicular distance between the surfaces. Between the cones is an assembly consisting of two balls within a housing which is movable parallel to the surface of the cones to maintain contact of the balls with the cones at the feet of the perpendicular beween the cones. It is assumed that pure rolling is maintained between the balls and the cones, and between themselves, and that there is frictionless sliding between the balls and the housing. With the assumption of negligible mass of the whole system the torques on the two shafts are proportional, with the proportionality being the instantaneous ratio of cone radii. The assumption of pure rolling means that the angular velocities are similarly proportional. Thus we may presume laws of the form:

$$T_1 = pT, \tag{6}$$

$$\omega_1 = -p^{-1}\omega \tag{7}$$

where T, T_1 are the torques on the shafts, ω, ω_1 are their angular velocities, and $p = p(t) > 0$ is the instantaneous ratio of cone radii. We note that $T_1\omega_1 + T\omega = 0$ so that no energy is absorbed or dissipated in the ideal device. Hence we may consider the schematic of Fig. 4 as a physical realisation of a lossless adjustable rotary transformer.

It is important to emphasize that, besides being lossless, an essential feature of the mechanism in Fig. 4 is that the ratio between angular velocities can be freely adjusted, including the case where the angular velocities are zero, as occurs when there is a reversal of sign. This feature contrasts with typical concepts of a continuously variable transmission (CVT), e.g., Brokowski et al. (2002); Rotella and Cammalleri (2018).

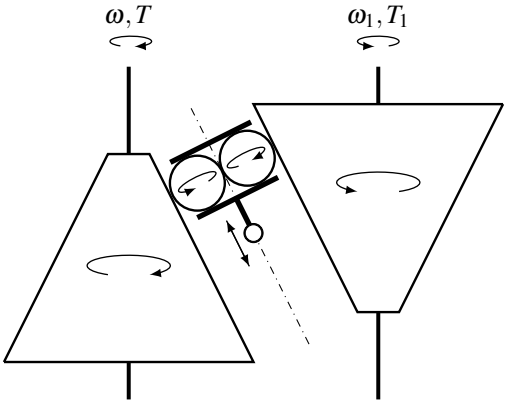


Fig. 4. Schematic of an adjustable rotary transformer with counter-rotating cones and continuously movable connecting assembly consisting of a pair of rotating balls within a housing.

2.6 The rotary varspring and varinerter

We first consider attaching a rotary spring of rotational stiffness $k > 0$ (constant) to the second shaft in Fig. 4 defined by $T_1 = -k\theta_1$ where $\dot{\theta}_1 = \omega_1$. A passive (lossless) rotary mechanical one-port is formed with the following relationship between the equal and opposite torque applied to the external (rotary) terminals T and the relative angular velocity ω between the terminals: $\omega = -p\omega_1 = pk^{-1}\frac{d}{dt}(pT)$. Similarly, if a rotary inerter (see Smith (2001)) with rotational inertance $b > 0$, defined by $T_1 = -b\dot{\omega}_1$, is connected across the second shaft in Fig. 4 a passive (lossless) rotary one-port is formed satisfying $T = -br\dot{\omega}_1 = br\frac{d}{dt}(r\omega)$. The constants k and b can be absorbed into p and r respectively (or equivalently setting $k = 1$ and $b = 1$).

This motivates the following definitions of the rotary varspring and varinerter:

$$\omega = p\frac{d}{dt}(pT) \qquad \text{(rotary varspring)} \tag{8}$$

$$T = r\frac{d}{dt}(r\omega) \qquad \text{(rotary varinerter)} \tag{9}$$

where (T, ω) is the torque-angular-velocity pair of the mechanical one-port and $p(t)$, $r(t)$ are positive and freely adjustable parameters. The internal energy of the devices is given by $\frac{1}{2}p^2T^2$ and $\frac{1}{2}r^2\omega^2$ respectively.

It is interesting to compare the embodiments presented for the translational and rotary varsprings and varinerters. A practical issue that arises with continuous operation of the translational devices, implemented in the manner of Fig. 3, is that the movement of the fulcrum in the x-direction may exceed the allowable travel. No such issue arises with the rotary devices.

3. PHYSICAL REALISATION OF A LOSSLESS ADJUSTABLE TWO-PORT TRANSFORMER

The conceptual embodiment of the adjustable rotary transformer in Fig. 4 encounters the technical challenge to ensure (nearly) frictionless sliding between the balls and the housing, while maintaining pure rolling between the two balls, and also between the balls and the cone surfaces. Next we describe a physical realisation for the housing and carriage assembly that offers a solution to this challenge.

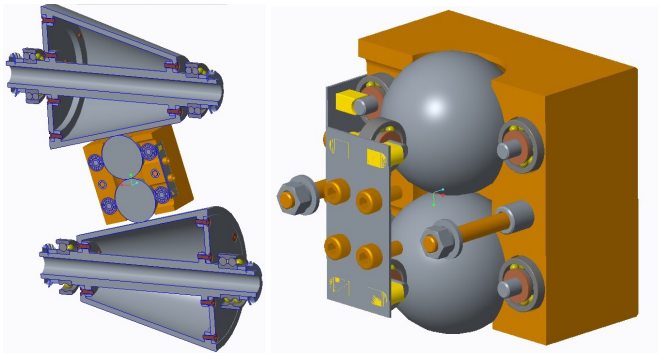


Fig. 5. Drawing of the device that consists of two shafts, two circular cones, and in between a carriage assembly with two counter-rotating balls (together with magnified view).

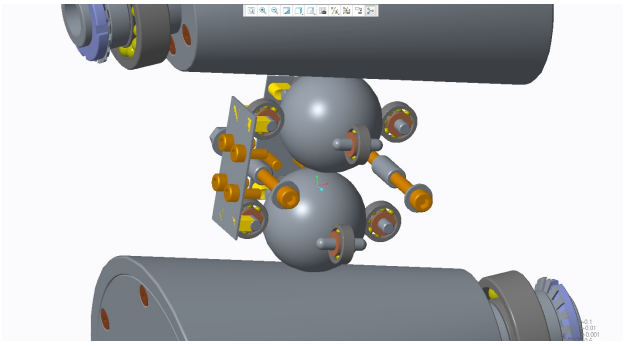


Fig. 6. Detail on the positioning of the counter-rotating balls with the supporting roller bearings.

The drawing in Fig. 5 shows the positioning of the two circular cones together with the housing that supports the two counter-rotating balls. Preloading of the shafts that support the cones can ensure sufficiently large normal forces at the contact points between the two balls, and between the balls and the cones, to ensure rolling (with no slippage) for a given specification on torques applied to the shafts. Preloading needs to ensure that the Hertzian contact stress is small enough so that the resulting strains are within the elastic limit. The housing encloses the two counter-rotating balls each constrained by two pairs of roller bearings positioned on perpendicular axes. Fig. 6 shows the arrangement of the roller bearings, each pair being on perpendicular axes which are themselves perpendicular to the line joining the centres of the two balls. This geometry allows each ball to freely rotate about any axis that lies within the plane whose normal is the line joining the centres of the two balls. This feature allows the carriage assembly to slide and be positioned along the axis (marked with dash-dotted line in Fig. 4) that runs parallel to the contact points between the balls and the two cones. This axis is not shown in Figs. 5–6 to allow a clear view of the internal structure of the housing and the positioning of the roller bearings. A mechanical embodiment of the device is shown in Fig. 7

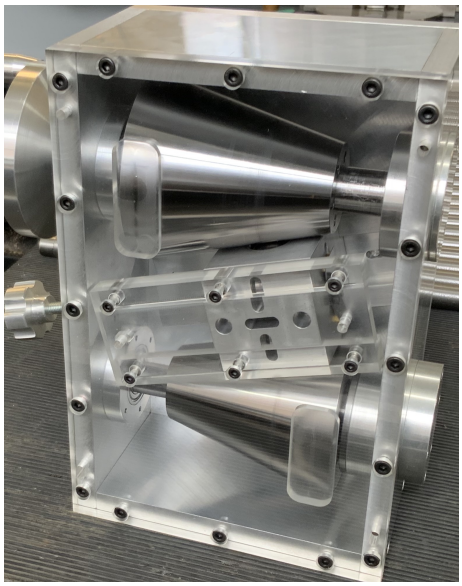


Fig. 7. Mechanical embodiment manufactured in Cambridge University Engineering Department.

With the assumption of negligible mass for the whole system, the torques on the two shafts are proportional, with the proportionality being the instantaneous ratio of cone radii. Further, no (or negligible) energy is absorbed or dissipated since any forces applied at the contact points are perpendicular to any direction that the contact points can be displaced.

4. CONCLUSION

We have shown how adjustable lossless springs and inerters can be realised. The central element of our construction is an adjustable mechanical transformer which is interesting in its own right. In this paper, we discuss the realisation of such a device and describe an embodiment that uses two counter-rotating cones that are coupled through two counter-rotating balls, housed in a carriage assembly that allows (near) friction-less motion of the balls while transfering torque between the shafts of the two cones.

REFERENCES

Bobrow, J.E., Jabbari, F., and Thai, K. (1995). An active truss element and control law for vibration suppression. *Smart Materials and Structures*, 4(4), 264.

Brezas, P., Smith, M.C., and Hoult, W. (2015). A clipped-optimal control algorithm for semi-active vehicle suspensions: Theory and experimental evaluation. *Automatica*, 53, 188–194.

Brokowski, M., Kim, S., Colgate, J.E., Gillespie, R.B., and Peshkin, M. (2002). Toward improved CVTs: Theoretical and experimental results. In *ASME 2002 International Mechanical Engineering Congress and Exposition*, 855–865. American Society of Mechanical Engineers.

Brzeski, P., Kapitaniak, T., and Perlikowski, P. (2015). Novel type of tuned mass damper with inerter which enables changes of inertance. *Journal of Sound and Vibration*, 349, 56–66.

Butsuen, T. and Hedrick, J. (1989). Optimal semi-active suspensions for automotive vehicles: The 1/4 car model. In *Advanced automotive technologies*, 305–319. ASME.

Chen, M.Z.Q., Hu, Y., Li, C., and Chen, G. (2014). Semi-active suspension with semi-active inerter and semi-active damper. *IFAC Proceedings Volumes*, 47(3), 11225–11230.

Garrido, H., Curadelli, O., and Ambrosini, D. (2018). On the assumed inherent stability of semi-active control systems. *Engineering Structures*, 159, 286–298.

Georgiou, T.T., Jabbari, F., and Smith, M.C. (2020). Principles of lossless adjustable one-ports. *IEEE Transactions on Automatic Control*, 65, 252–262.

Jabbari, F. and Bobrow, J.E. (2002). Vibration suppression with resettable device. *Journal of Engineering Mechanics*, 128(9), 916–924.

Lazarek, M., Brzeski, P., and Perlikowski, P. (2018). Design and identification of parameters of tuned mass damper with inerter which enables changes of inertance. *Mechanism and Machine Theory*, 119, 161–173.

Rotella, D. and Cammalleri, M. (2018). Power losses in power-split CVTs: A fast black-box approximate method. *Mechanism and Machine Theory*, 128, 528–543.

Savaresi, S.M., Poussot-Vassal, C., Spelta, C., Sename, O., and Dugard, L. (2010). *Semi-active suspension control design for vehicles*. Elsevier.

Smith, M.C. (2002). Synthesis of mechanical networks: The inerter. *IEEE Transactions on automatic control*, 47(10), 1648–1662.

Smith, M.C. (2020). The inerter: a retrospective. *Annual Review of Control, Robotics, and Autonomous Systems*, 3, 361–391.

Smith, M.C., Hoult, W., and Brezas, P. (2018). McLaren earns its Ph.D in handling. *Automotive Engineering*, 5(7), 34–35.

Smith, M.C. (2001). Force-controlling mechanical device. US Patent 7,316,303.

Vanderborght, B., Albu-Schäffer, A., Bicchi, A., Burdet, E., Caldwell, D.G., Carloni, R., Catalano, M., Eiberger, O., Friedl, W., Ganesh, G., et al. (2013). Variable impedance actuators: A review. *Robotics and autonomous systems*, 61(12), 1601–1614.

Wolf, S., Grioli, G., Eiberger, O., Friedl, W., Grebenstein, M., Höppner, H., Burdet, E., Caldwell, D.G., Carloni, R., Catalano, M.G., et al. (2016). Variable stiffness actuators: Review on design and components. *IEEE/ASME transactions on mechatronics*, 21(5), 2418–2430.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Route planning and hybrid games for multiple players ⋆

S. Cacace * **A. Festa** ** **R. Ferretti** ***

* *Dipartimento di Matematica e Fisica, Università Roma Tre, Roma, Italy (e-mail: cacace@mat.uniroma3.it)*
** *Dipartimento di Scienze Matematiche "Giuseppe Luigi Lagrange", Politecnico di Torino, Torino, Italy (e-mail: adriano.festa@polito.it)*
*** *Dipartimento di Matematica e Fisica, Università Roma Tre, Roma, Italy (e-mail: ferretti@mat.uniroma3.it)*

Abstract: We investigate the modelling of sailing races as hybrid stochastic games, either with zero or nonzero sum, where the first case is typical of match races and the second of fleet races. In particular, we provide models of growing complexity and dimension, study the optimal strategies in various racing situations and devise some fast and/or reduced memory implementation.

Keywords: Hybrid Systems, Optimal Control, Stochastic Control and Estimation

1. INTRODUCTION

In its typical formulation, the *route planning* problem consists in driving a sailing vessel towards a target in a partly stochastic wind field. This problem has recently received a certain attention, not only because of its use in sailing competitions, but also for the growing interest in sustainable transport strategies. Actually, various projects of sailing or hybrid commercial ships are currently under consideration, and their use could significantly reduce costs and emissions related to the transport on sea: exploiting in an optimal way the wind field would thus become a crucial point for such projects. In sailing races, the problems of optimizing the boat route appears critical, and has already motivated a certain amount of studies, in particular related to America's Cup competitions (see Philpott (2005); Vinckenbosch (2012)).

The main peculiarity in modelling the motion of sailing boats is that their speed depends on the angle between the boat direction and the (random) direction of the wind, and presents a *no-go* region if the boat attempts to sail opposite to the wind. A second point is that changing from right to left tack of vice versa (i.e., changing the side of the boat from which the wind comes) requires a complex manoeuvre (*tacking* or *gybing*) which causes a speed loss, to be correctly taken into account. A typical polar plot of speeds is shown in Fig. 1, in which the arrows refer to the speed obtained at various angles, on the left tack (i.e., with the wind arriving from left).

The interpretation of the tacking or gybing manoeuvre as a *switch* between two different dynamics allows us to formulate the problem in terms of *hybrid control* (see Ferretti and Festa (2019)), where the angle with the wind on a given tack is considered as a continuous control, while a tacking or gybing represents a commutation between the dynamics associated to respectively right or left tack.

⋆ This research has been partially supported by INdAM–GNCS and by the MIUR grant "Dipartimenti Eccellenza 2018–2022".

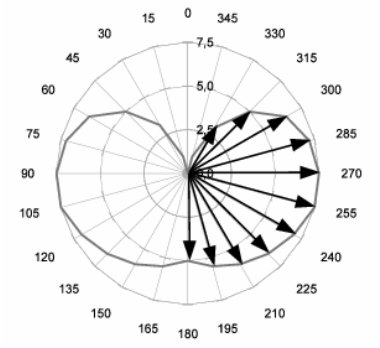


Fig. 1. Polar plot of the speed of a sailing boat with respect to the wind angle

A time loss (in other terms, a cost) is associated to this change of dynamics.

2. MODELLING A RACE

In this work, we will apply this general hybrid stochastic framework to the case of a sailing race, in which a set of players (boats) $A, B, \dots$ are placed in a two-dimensional domain representing the race area. The wind direction $\Theta(t)$ will be modelled, as usual in this framework, by adding a deterministic drift γ to a Brownian motion, so that

$$d\Theta(t) = \gamma dt + \sigma dW(t),$$

while, on the X_1–X_2 plane, the generic sailing boat P has a position evolving as

$$\begin{cases} \dot{X}_1^P(t) = s^P(a^P, X^A, X^B, \dots) \sin\left(\Theta(t) + (-1)^{Q^P(t)} a^P\right) \\ \dot{X}_2^P(t) = s^P(a^P, X^A, X^B, \dots) \cos\left(\Theta(t) + (-1)^{Q^P(t)} a^P\right), \end{cases}$$

where a is the angle of the boat w.r.t. the wind, $s^P(a^P, X^A, X^B, \dots)$ is the boat speed, and the discrete control $Q^P(t) = 1$ for the port tack, $Q^P(t) = 2$ on the

starboard tack. Note that the boat speed depends on the positions of all the players, thus including the effects of mutual disturbance, which occurs in real races (and that represent a key feature of the problem).

2.1 Various settings for the problem

We will consider various settings of increasing complexity, all of them being of hybrid stochastic game type, for a race among players of the form above. The basic problem is sailing to windward in an infinite plane, as a model for windward strategy when far from the target. This setting has been used in Cacace et al. (2020) for treating the case of match races, and in this situation we consider a zero-sum game of pursuit–evasion form, which can be conveniently treated in reduced coordinates with a state space of dimension $n = 3$. Using a dynamic programming approach, the value function v of the game solves an Isaacs equation in the form of a system of quasi-variational inequalities:

$$\max\{v - \mathcal{M}[v], \min\{v - \mathcal{N}[v], F[v]\}\} = 0,$$

where $F[v]$ is the dynamic programming Hamiltonian associated to the continuous controls, and $\mathcal{M}$, $\mathcal{N}$ are the switching operators of the two players. The numerical examples carried out in Cacace et al. (2020) show that the optimal strategy for both players is basically to follow the optimal single player strategy, but as one of the players gains some advantage, this player tends to disturb slightly the other one, whenever in favourable position (see Figures 2–3).

In this work, we will examine various generalizations of this setting, and in particular:

(1) Sailing to windward in an infinite plane, but in a nonzero-sum framework. In this case we expect that a nontrivial Nash equilibrium may appear, in which the players are not interested to disturb each other. This may be a first model for the onset of Nash equilibria in a fleet race. However, the case can no longer be treated in reduced coordinates, and this makes the dimension of the state space raise to $n = 2N + 1$, where N is the number of players;
(2) Sailing to windward, either in a zero- or nonzero-sum game, with a target. In this case, while the nonzero-sum game appears as a multiple minimum time problem, in the zero-sum game we need to define a suitable criterion of advantage for a player. The dimension of the state space is still $n = 2N + 1$, the use of reduced coordinates being impossible.

2.2 Computational issues

As we have outlined above, the state space may rise to a relatively high dimension, as soon as the model becomes realistic. In addition, a slow convergence may occur if the numerical approximation is implemented in the form of a plain value iteration. We will therefore discuss:

(1) A memory reduction technique, which uses the fact that the interplay between two players appears only at small distances, and therefore the coupled game reduces to an optimal control problem for a single player as soon as the players are far enough from each other;
(2) Solvers of *fast marching* or *fast sweeping* type, exploiting the direction of propagation of the value function, along with the degeneracy of the stochastic component of the dynamics;
(3) Solvers of *policy iteration* type.

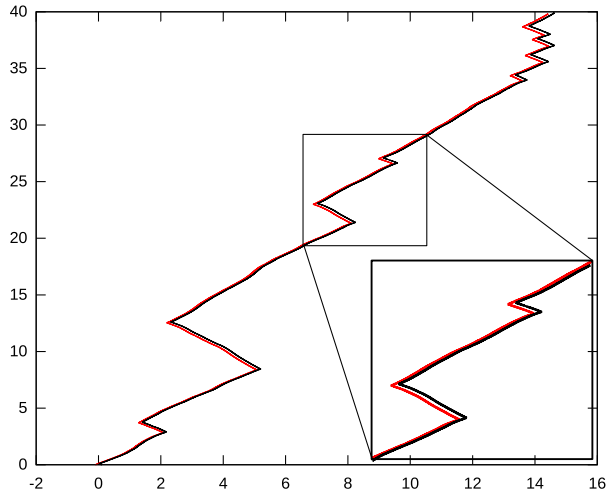


Fig. 2. A sample trajectory of the zero-sum game in symmetric conditions, with the black player leading

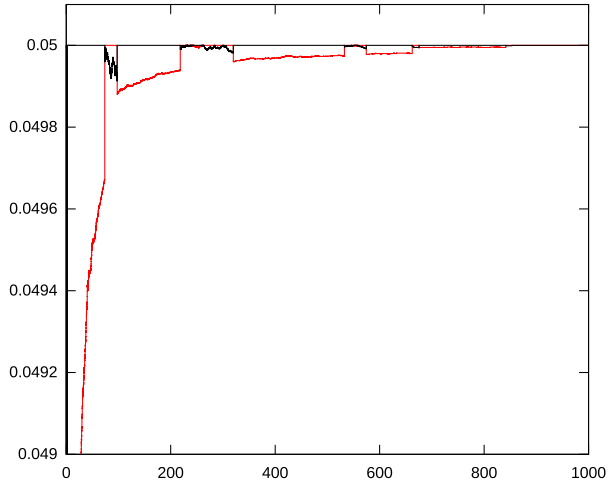


Fig. 3. Speeds of the two players for the example above

REFERENCES

Cacace, S., Ferretti, R., and Festa, A. (2020). Stochastic hybrid differential games and match race problems. *Applied Mathematics and Computation*, 372, 124966.

Ferretti, R. and Festa, A. (2019). Optimal route planning for sailing boats: A hybrid formulation. *Journal of Optimization Theory and Applications*, 181(3), 1015–1032.

Philpott, A. (2005). Stochastic optimization and yacht racing. In *Applications of stochastic programming*, 315–336. SIAM.

Vinckenbosch, L. (2012). *Stochastic control and free boundary problems for sailboat trajectory optimization.* PhD thesis, EPFL, Lausanne.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On the Optimal Control of Lossless Electrical Networks

Richard Pates [1]

Abstract: Electrical networks constructed out of resistors (R), inductors (L), capacitors (C), transformers (T), and gyrators (G) are used throughout engineering and the applied sciences to model physical processes. Synthesising RLCTG networks for control purposes is also important, since in a number application domains the corresponding controllers can be implemented without an energy source. We show that if a process can be modelled by an LCTG network, a controller that maximises robustness with respect to normalised coprime factor perturbations can be synthesised by a decentralised resistive network. The results are illustrated on an example centred on the iterative solution to constrained least squares problems.

Keywords: Dissipativity, Robust and H-infinity control, Large-scale systems

NOTATION

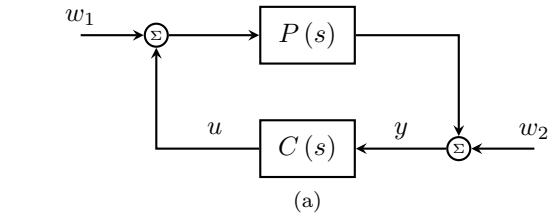
$\mathcal{L}_2^n$ and $\mathcal{L}_2^{\mathrm{loc},n}$ denote the n-vectors of square integrable functions, and locally square integrable functions, respectively. $\mathbb{R}^{n\times m}[s]$ denotes the n by m matrices of polynomials in the indeterminate s with real coefficients. $\|\cdot\|$ denotes either the $\mathcal{L}_2$ norm or the matrix 2-norm depending on whether it acts on a function or matrix. I denotes the identity matrix.

1. INTRODUCTION

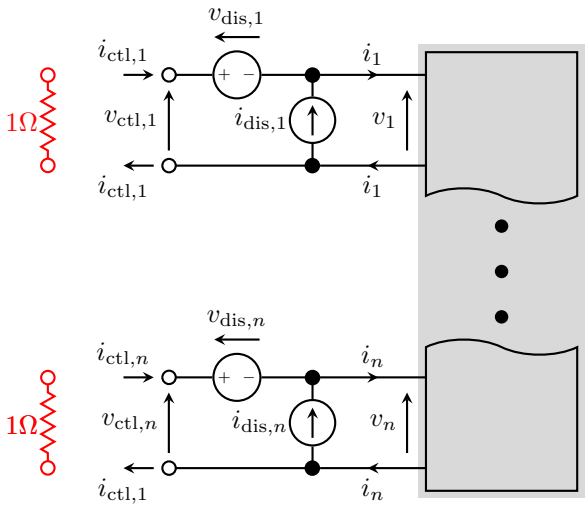
We investigate the H-infinity control problem that underpins the loop-shaping design procedure of McFarlane and Glover (1992) in the context of electrical networks. Our main contribution is to show that when the process to be controlled corresponds to an LCTG network (an electrical network constructed only from inductors, capacitors, transformers and gyrators), the resulting optimal control law is decentralised, and can be implemented by connecting unit resistors across all the external terminal pairs of the network. This is illustrated in Fig. 1. This gives an example of a class of optimal control problems with inherently decentralised solutions, providing motivation for applying decentralised control in applications with lossless (or nearly lossless) dynamics, such as electrical power systems.

As discussed in Camlibel et al. (2003), the dynamics of electrical networks cannot always be described in the input-output framework. However the driving point currents $i \in \mathcal{L}_2^{\mathrm{loc},n}$ and volatages $v \in \mathcal{L}_2^{\mathrm{loc},n}$ of an electrical network constructed out of resistors, inductors, capacitors, transformers and gyrators can always be characterised by the solutions to an equation that takes the form

$$R\left(\tfrac{d}{dt}\right)\begin{bmatrix} i \\ v \end{bmatrix} = 0,$$



(a)



(b)

Fig. 1. (a) The standard setup for the H-infinity control problem that underpins the loop-shaping design procedure of McFarlane and Glover (1992). The objective is to minimise the H-infinity norm of the transfer function from $[w_1^\mathsf{T}\, w_2^\mathsf{T}]^\mathsf{T}$ to $[u^\mathsf{T}\, y^\mathsf{T}]^\mathsf{T}$. (b) Electrical equivalent of the H-infinity control problem in (a). The main contribution of the paper is to show that if the electrical network is an LCTG network (it is constructed using only inductors, capacitors, transformers and gyrators), then this H-infinity performance criterion is minimised by connecting unit resistors across all the open terminal pairs.

[1] The author is a member of the ELLIIT Strategic Research Area at Lund University. This work was supported by the ELLIIT Strategic Research Area. This project has received funding from VR grant 2016-04764 and ERC grant agreement No 834142.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

where $R(s) \in \mathbb{R}^{n\times 2n}[s]$ (see Hughes (2017b)). We therefore study the following behavioral equivalent of the H-infinity control problem in Fig. 1(a).

Problem 1: Let $G(s) \in \mathbb{R}^{\cdot\times m}[s]$ define the uncontrolled behavior of a process according to

$$\mathfrak{B}_{G(s)} = \left\{(w,z) \in \mathcal{L}_2^{\mathrm{loc},m} \times \mathcal{L}_2^{\mathrm{loc},m} : G\left(\tfrac{d}{dt}\right)(w+z) = 0\right\}.$$

Choose $K(s) \in \mathbb{R}^{\cdot\times m}[s]$ to minimise

$$\sup\left\{\|z\| : w \in \mathcal{L}_2^m, \|w\| = 1, (w,z) \in \mathfrak{B}_{G(s),K(s)}\right\} \quad (1)$$

where $\mathfrak{B}_{G(s),K(s)} = \left\{(w,z) \in \mathfrak{B}_{G(s)} : K\left(\tfrac{d}{dt}\right) z = 0\right\}$.

To see the connection, partition R as

$$R(s) = \left[R_1(s) \ -R_2(s)\right],$$

and suppose that

$$R\left(\tfrac{d}{dt}\right) v = 0 \quad \Leftrightarrow \quad \left[R_2\left(\tfrac{d}{dt}\right)^{-1} R_1\left(\tfrac{d}{dt}\right) \ -I\right] v = 0.$$

That is the behavior associated with $R(s)$ admits an input-output representation with respect to the given partition, with transfer function

$$P(s) = R_2(s)^{-1} R_1(s).$$

Splitting w and z compatibly according to

$$w = \begin{bmatrix} w_1 \\ -w_2 \end{bmatrix} \text{ and } z = \begin{bmatrix} u \\ y \end{bmatrix},$$

we then see that $(w,z) \in \mathfrak{B}_{G(s)}$ if and only if

$$y = w_2 + P\left(\tfrac{d}{dt}\right)(u + w_1),$$

placing Problem 1 in the standard H-infinity control framework. Problem 1 therefore represents a generalisation to cases where this input-output representation is not necessarily possible.

In the context of this paper, the idea behind Problem 1 is that $G(s)$ characterises the dynamics of an electrical network, w a set of disturbance currents and voltages, and z a set of driving point currents and voltages that can be constrained by another electrical network:

$$w = \begin{bmatrix} i_{\mathrm{dis}} \\ -v_{\mathrm{dis}} \end{bmatrix} \text{ and } z = \begin{bmatrix} i_{\mathrm{ctl}} \\ v_{\mathrm{ctl}} \end{bmatrix}.$$

Since Kirchhoff's current and voltage laws at the terminal pairs of the electrical network read as

$$\begin{bmatrix} i \\ v \end{bmatrix} = \begin{bmatrix} i_{\mathrm{ctl}} \\ v_{\mathrm{ctl}} \end{bmatrix} + \begin{bmatrix} i_{\mathrm{dis}} \\ -v_{\mathrm{dis}} \end{bmatrix},$$

the objective is then to design an electrical network that can be connected to the given electrical network to minimise the effect of the disturbance currents and voltages as quantified by (1). This particular performance objective has a wide range of interpretations (see §2 of Vinnicombe (2000) for an extended discussion), including maximising robustness to normalised coprime factor perturbations of the process dynamics.

The rest of the paper is structured as follows. We begin by demonstrating that whenever the process to be controlled corresponds to an LCTG network, the optimal solution to Problem 1 is given by

$$K(s) \equiv [I \ I].$$

This control law can be implemented by connecting a unit resistor across each of the open terminal pairs of the perturbed electrical network, as depicted in Fig. 1(b). This is the implication (1) ⇒ (2) in Theorem 1. Theorem 1 also demonstrates that the LCTG networks can be characterised in terms of solutions to Problem 1. This is more of a curiosity than anything else, but offers an alternative classification of the LCTG networks to the conditions on p.21 of Hughes and Branford (2021). The paper concludes with an example showing that this optimal control law arises naturally in a simple iterative algorithm for solving constrained least squares problems.

2. RESULTS

Theorem 1. Given $R(s) \in \mathbb{R}^{n\times 2n}[s]$, the following are equivalent:

(1) The behavior

$$\mathfrak{B} = \left\{\begin{bmatrix} i \\ v \end{bmatrix} \in \mathcal{L}_2^{\mathrm{loc},2n} : R\left(\tfrac{d}{dt}\right)\begin{bmatrix} i \\ v \end{bmatrix} = 0\right\}$$

is the driving point behavior of an LCTG network, where i and v denote the driving point currents and voltages, respectively.

(2) The matrix

$$K(s) \equiv [I \ I]$$

minimises the performance objective in Problem 1 when $G(s) \equiv R(s)$, and

$$K(s) \equiv [I \ -I]$$

minimises the performance objective in Problem 1 when $G(s) \equiv R(-s)$. In both cases a value of $\sqrt{2}$ is achieved for the performance objective in (1).

Proof. (1) ⇒ (2): It follows from Theorem 5 of Hughes (2017a) that under the hypothesis that $\mathfrak{B}$ is the driving point behavior of an LCTG network, there exists a permutation matrix $P = [P_1 \ P_2]$, and matrices

$$\begin{bmatrix} -A & -B \\ C & D \end{bmatrix} = -\begin{bmatrix} -A & -B \\ C & D \end{bmatrix}^{\mathsf{T}}, \quad (2)$$

where the pair (C,A) is observable and (A,B) is controllable [2], such that

$$\mathfrak{B} = \left\{\begin{bmatrix} i \\ v \end{bmatrix} : \left[\begin{array}{c} i \\ \hdashline v \end{array}\right] \left[\begin{array}{cc:cc} P_1 & 0 & P_2 & 0 \\ \hdashline 0 & P_2 & 0 & P_1 \end{array}\right] \left[\begin{array}{c} u \\ \hdashline y \end{array}\right], (u,y) \in \mathfrak{B}_{\mathrm{ss}}\right\},$$

where

$$\mathfrak{B}_{\mathrm{ss}} = \left\{(u,y) : \begin{bmatrix} x \\ u \\ y \end{bmatrix} \in \mathcal{L}_2^{\mathrm{loc},2n+m}, \begin{bmatrix} A - I\frac{d}{dt} & B & 0 \\ C & D & -I \end{bmatrix}\begin{bmatrix} x \\ u \\ y \end{bmatrix} = 0\right\}.$$

It then follows that $\mathfrak{B}_{R(s)}$ is given by the set of weak solutions to the equations

$$\begin{aligned} \tfrac{d}{dt}x &= Ax + [B \ 0]\,d + Bu, \\ F^{\mathsf{T}}z &= \left(\begin{bmatrix} C \\ 0 \end{bmatrix}x + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}d + \begin{bmatrix} D \\ I \end{bmatrix}u\right), \\ y &= Cx + [0 \ I]\,d + Du, \end{aligned} \quad (3)$$

where

$$F = \begin{bmatrix} P_1 & 0 & P_2 & 0 \\ 0 & P_2 & 0 & P_1 \end{bmatrix} \text{ and } w = Fd.$$

Since $FF^{\mathsf{T}} = F^{\mathsf{T}}F = I$ and the realisation in (3) is minimal, Problem 1 with $G(s) \equiv R(s)$ corresponds to minimising the H-infinity norm of the transfer function

[2] The existence of an observabile and controllabile realisation can be deduced from Hughes (2017b), since together Definition 7, Remark 8 and Theorem 9 from that paper imply that the driving point behavior of an LCTG network is always behaviorally controllable. See also the proof of Theorem 2 from Pates (2022) for an approach based on state-space techniques.

from d to $F^{\mathsf{T}}z$ as defined by (3). This is precisely the H-infinity control problem studied in Glover and McFarlane (1989). If we let

$$R = \left(I + DD^{\mathsf{T}}\right) \text{ and } S = \left(I + D^{\mathsf{T}}D\right),$$

the optimal value for this problem is equal to

$$\sqrt{1 + \lambda_{\max}},$$

where $\lambda_{\max}$ is the largest eigenvalue of XY, X is the unique positive semi-definite solution to the generalised control algebraic Riccati equation

$$\begin{aligned}\left(A - BS^{-1}D^{\mathsf{T}}C\right)^{\mathsf{T}} X + X\left(A - BS^{-1}D^{\mathsf{T}}C\right)\\ -XBS^{-1}B^{\mathsf{T}}X + C^{\mathsf{T}}R^{-1}C = 0,\end{aligned}$$

and Z is the unique positive semi-definite solution to the generalised filtering algebraic Riccati equation

$$\begin{aligned}\left(A - BS^{-1}D^{\mathsf{T}}C\right)Z + Z\left(A - BS^{-1}D^{\mathsf{T}}C\right)^{\mathsf{T}}\\ -ZC^{\mathsf{T}}R^{-1}CZ + BS^{-1}B^{\mathsf{T}} = 0.\end{aligned}$$

From (2) we see that $A = -A^{\mathsf{T}}$, $B = C^{\mathsf{T}}$, and $D = -D^{\mathsf{T}}$. Direct substitution then reveals that $X \equiv I$ and $Z \equiv I$ solve the Riccati equations under these constraints. It is easily checked that the optimal value is achieved by the control law $u = -y$, which corresponds to

$$K(s) \equiv [I \ I].$$

A state-space description for $\mathfrak{B}_{R(-s)}$ can be similarly obtained (make the substitutions $A \mapsto -A$ and $B \mapsto -B$ in (3)), leading once more to the Riccati equation solutions $X \equiv I$ and $Z \equiv I$, but this time yielding the optimal control law defined by

$$K(s) \equiv [I \ -I]$$

as required.

(2) $\Rightarrow$ (1): First observe that for any $s \in \mathbb{C}$,

$$\operatorname{rank} R(s)\begin{bmatrix} I \\ -I \end{bmatrix} \leq \min\{\operatorname{rank} R(s), n\}$$

and so there exist no non-zero $v \in \mathcal{L}_2^{\mathrm{loc},n}$ such that

$$R\left(\tfrac{d}{dt}\right)\begin{bmatrix} I \\ -I \end{bmatrix} v = 0$$

only if (Polderman and Willems, 1998, Lemma 5.4.8)

$$\operatorname{rank} R(s) = n \text{ for all } s \in \mathbb{C}. \tag{4}$$

Next observe that if $G(s) \equiv R(s)$ and $K(s) \equiv [I \ I]$, then $\mathfrak{B}_{G(s),K(s)}$ is equal to the set of locally integrable (w, z) such that

$$R\left(\tfrac{d}{dt}\right)\left(w + \begin{bmatrix} I \\ -I \end{bmatrix} v\right) = 0 \text{ and } z = \begin{bmatrix} I \\ -I \end{bmatrix} v.$$

for some $v \in \mathcal{L}_2^{\mathrm{loc},n}$. Therefore under the hypothesis of (2), (4) must hold, otherwise there would exist a v with $\|v\| = 1/\sqrt{2}$ such that for any $\alpha \in \mathbb{R}$

$$\left(-\begin{bmatrix} I \\ -I \end{bmatrix} v, (1+\alpha)\begin{bmatrix} I \\ -I \end{bmatrix} v\right) \in \mathfrak{B}_{R(s),K(s)},$$

which would imply that the performance objective in (1) is infinite. This implies that $\mathfrak{B}$ is behaviorally controllable, from which it follows (Willems, 1991, §8) that there exist matrices of rational functions $\tilde{M}(s)$ and $\tilde{N}(s)$ with no poles in the closed right-half-plane such that

$$\begin{bmatrix} i \\ v \end{bmatrix} \in \mathfrak{B} \quad \Leftrightarrow \quad [-\tilde{M}\left(\tfrac{d}{dt}\right) \ \tilde{N}\left(\tfrac{d}{dt}\right)]\begin{bmatrix} i \\ v \end{bmatrix} = 0.$$

Furthermore it is no loss of generality to assume that this left coprime factorisation is normalised (Vidyasagar (1988)), meaning that for all $\omega \in \mathbb{R}$

$$\tilde{M}(j\omega)\,\tilde{M}(j\omega)^* + \tilde{N}(j\omega)\,\tilde{N}(j\omega)^* = I. \tag{5}$$

From the inequalities on p.70 of Vinnicombe (2000), for this control law to achieve a performance value of $\sqrt{2}$,

$$\sup\left\{\left\|[-\tilde{M}(s) \ \tilde{N}(s)]\begin{bmatrix} \frac{1}{\sqrt{2}}I \\ -\frac{1}{\sqrt{2}}I \end{bmatrix}\right\| : s \in \mathbb{C}, \operatorname{Re}(s) > 0\right\} \geq \frac{1}{\sqrt{2}}.$$

Using (5), the above is equivalent to the inequality

$$\tilde{N}(s)\,\tilde{M}(s)^* + \tilde{M}(s)\,\tilde{N}(s)^* \succeq 0 \tag{6}$$

holding for all $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 0$. Repeating the above steps for $G(s) \equiv R(-s)$ and $K(s) \equiv [I \ -I]$ demonstrates that[3] for all $\omega \in \mathbb{R}$,

$$\tilde{N}(j\omega)\,\tilde{M}(j\omega)^* + \tilde{M}(j\omega)\,\tilde{N}(j\omega)^* \preceq 0. \tag{7}$$

It then follows from (4), (6) and (7) that all the conditions from (Hughes and Branford, 2021, p.21) are satisfied, making $\mathfrak{B}$ the driving point behavior of an LCTG network. □

3. EXAMPLE

The objective of constrained least squares is to find an $\bar{x} \in \mathbb{R}^n$ that satisfies

$$\min_{\bar{x}\in\mathbb{R}^n} \|C\bar{x} - b\|_2, \text{s.t. } A\bar{x} = d, \tag{8}$$

where $C \in \mathbb{R}^{m\times n}$, $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^m$ and $d \in \mathbb{R}^p$ are the problem data. Constrained least squares encompasses a very broad class of problems including, for example, finite horizon LQR, and includes standard least squares and minimum norm solutions to a set of linear equations as special cases ($p = 0$, and $C = I$ and $b = 0$, respectively). The solution to (8) can be obtained from the Karush-Kuhn-Tucker conditions

$$\begin{bmatrix} -C^{\mathsf{T}}C & -A^{\mathsf{T}} \\ A & 0 \end{bmatrix}\begin{bmatrix} \bar{x} \\ \bar{z} \end{bmatrix} = \begin{bmatrix} -C^{\mathsf{T}}b \\ d \end{bmatrix}, \tag{9}$$

where we assume for simplicity that

i) A is right invertible;
ii) $\begin{bmatrix} C \\ A \end{bmatrix}$ is left invertible;

so that these equations have a unique solution. We will now see that the solution to (9) arises naturally from the solution to Problem 1.

To this end, consider the system

$$\begin{aligned}\tfrac{d}{dt}x &= \begin{bmatrix} 0 & -A^{\mathsf{T}} \\ A & 0 \end{bmatrix}x + \begin{bmatrix} C^{\mathsf{T}} \\ 0 \end{bmatrix}(u + w_1 - r_1) + \begin{bmatrix} 0 \\ I \end{bmatrix}r_2,\\ y &= [C \ 0]x + w_2.\end{aligned} \tag{10}$$

Together (i)-(ii) imply that this realisation is observable and controllable, which implies (Willems, 1972, Theorem 8) that it can be synthesised using p inductors, n capacitors and some transformers. In light of Theorem 1, the control law that minimises (1) (with $r_1 = r_2 = 0$) can be synthesised by connecting unit resistors across each of the external terminal pairs. This maximises robustness with

[3] A small technical point here is that $\tilde{M}(-s)$ and $\tilde{N}(-s)$ do not give a left coprime factorization of $R(-s)$, since the resulting factors have poles in the right-half-plane. Nevertheless it is necessary that the resulting inequalities must still hold on the imaginary axis, which is all that is required to conclude (7).

respect to normalised coprime factor perturbations. We also see from (10) that the closed loop system becomes

$$\frac{d}{dt}x = \begin{bmatrix} -C^{\mathsf{T}}C & -A^{\mathsf{T}} \\ A & 0 \end{bmatrix} x + \begin{bmatrix} C^{\mathsf{T}} \\ 0 \end{bmatrix} (w_1 - r_1) + \begin{bmatrix} 0 \\ I \end{bmatrix} r_2,$$
$$y = [C\ 0]x + w_2.$$

Therefore by applying the step inputs $r_1 = bH(t)$ and $r_2 = dH(t)$, where $H(t)$ denotes the unit step, we see that

$$\lim_{t\to\infty} x(t) = \begin{bmatrix} \bar{x} \\ \bar{z} \end{bmatrix}.$$

This means that for any b and d, the solution to the constrained least squares problem can be obtained by measuring the voltage across the capacitors. Of course it is not necessary to actually use an electrical circuit to implement the optimal control law. These equations could equally well be interpreted as a simple algorithm which can be easily distributed in the case of sparse C and A, and Theorem 1 then shows that this algorithm has excellent robustness properties.

REFERENCES

Camlibel, M.K., Willems, J.C., and Belur, M.N. (2003). On the dissipativity of uncontrollable systems. In *42nd IEEE Conference on Decision and Control*, volume 2, 1645–1650. doi:10.1109/CDC.2003.1272848.

Glover, K. and McFarlane, D. (1989). Robust stabilization of normalized coprime factor plant descriptions with H_∞-bounded uncertainty. *IEEE Transactions on Automatic Control*, 34(8), 821–830. doi:10.1109/9.29424.

Hughes, T.H. (2017a). Passivity and electric circuits: a behavioral approach. *IFAC-PapersOnLine*, 50(1), 15500–15505. doi:10.1016/j.ifacol.2017.08.2117. 20th IFAC World Congress.

Hughes, T.H. (2017b). A theory of passive linear systems with no assumptions. *Automatica*, 86, 87–97. doi:10.1016/j.automatica.2017.08.017.

Hughes, T.H. and Branford, E.H. (2021). Dissipativity, reciprocity and passive network synthesis: from Jan Willems' seminal dissipative dynamical systems papers to the present day. *arXiv:2102.08855[math.OC]*.

McFarlane, D. and Glover, K. (1992). A loop-shaping design procedure using H_∞ synthesis. *IEEE Transactions on Automatic Control*, 37(6), 759–769. doi:10.1109/9.256330.

Pates, R. (2022). Passive and reciprocal networks: From simple models to simple optimal controllers. *arXiv:2201.12228[math.OC]*.

Polderman, J.W. and Willems, J.C. (1998). *Introduction to Mathematical Systems Theory: A Behavioral Approach*. 26. Springer Science & Business Media.

Vidyasagar, M. (1988). Normalized coprime factorizations for non-strictly proper systems. *IEEE Transactions on Automatic Control*, 33(3), 300–301. doi:10.1109/9.408.

Vinnicombe, G. (2000). *Uncertainty and Feedback, H_∞ Loop-Shaping and the ν-Gap Metric*. World Scientific Publishing Company.

Willems, J.C. (1991). Paradigms and puzzles in the theory of dynamical systems. *IEEE Transactions on Automatic Control*, 36(3), 259–294. doi:10.1109/9.73561.

Willems, J.C. (1972). Dissipative dynamical systems, Part II: Linear systems with quadratic supply rates. *Archive for rational mechanics and analysis*, 45(5), 321–351. doi:10.1007/BF00276494.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Tractable semidefinite bounds of positive maximal singular values ⋆

Victor Magron * **Ngoc Hoang Anh Mai** ** **Yoshio Ebihara** ***
Hayato Waki ****

* *LAAS-CNRS, Université de Toulouse, CNRS, IMT, Toulouse, France (e-mail: vmagron@laas.fr).*
** *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France (e-mail: nhmai@laas.fr)*
*** *Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka819-0395, Japan (e-mail: ebihara@ees.kyushu-u.ac.jp)*
**** *Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka819-0395, Japan (e-mail: waki@imi.kyushu-u.ac.jp)*

Abstract: We focus on computing certified upper bounds for the positive maximal singular value (PMSV) of a given matrix. The PMSV problem boils down to maximizing a quadratic polynomial on the intersection of the unit sphere and the nonnegative orthant. We provide a hierarchy of tractable semidefinite relaxations to approximate the value of the latter polynomial optimization problem as closely as desired. This hierarchy is based on an extension of Pólya's representation theorem. Doing so, positive polynomials can be decomposed as weighted sums of squares of s-nomials, where s can be a priori fixed ($s = 1$ corresponds to monomials, $s = 2$ corresponds to binomials, etc.). This in turn allows us to control the size of the resulting semidefinite relaxations.

Keywords: Polynomial optimization, Pólya's representation, semidefinite programming, linear programming, second-order conic programming, s-nomials, positive maximal singular value
AMS subject classifications: 90C22, 90C26

1. INTRODUCTION

Let $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+$ be the sets of nonnegative integers, real numbers and nonnegative real numbers, respectively. In the present paper, we focus on the problem of computing the positive maximal singular value (PMSV) of a given real matrix $M \in \mathbb{R}^{n\times n}$, denoted by $\sigma_+(M)$. Providing PMSV bounds is crucial for certain induced norm analysis of discrete-time linear time-invariant systems, where the input signals are restricted to be nonnegative; see, e.g., Ebihara et al. (2021). The squared value $\sigma_+(M)^2$ can be obtained by solving the following polynomial maximization problem:

$$\begin{aligned}\sigma_+(M)^2 &= \sup_{x\in\mathbb{R}^n_+}\{x^\top(M^\top M)x \,:\, \|x\|_2^2 = 1\}\,,\\ &= \sup_{x\in\mathbb{R}^n_+}\{x^\top(M^\top M)x \,:\, \|x\|_2^2 \le 1\}\,.\end{aligned} \tag{1}$$

For $m \in \mathbb{N}\backslash\{0\}$, let $[m] := \{1,\dots,m\}$. Note that (1) is a particular instance of a polynomial optimization problem (POP) on the nonnegative orthant:

$$f^\star := \sup_{x\in S} f(x)\,, \tag{2}$$

where f is a polynomial and S is a basic closed semialgebraic set, i.e., the intersection of finitely many polynomial inequalities as follows:

$$S := \{x \in \mathbb{R}^n \,:\, x_i \ge 0\,,\, i \in [n]\,,\, g_j(x) \ge 0\,,\, j \in [m]\}\,, \tag{3}$$

for some $g_j \in \mathbb{R}[x]$, $j \in [m]$ with $g_m := 1$. Letting $x^2 := (x_1^2,\dots,x_n^2)$ and $\check{p}(x) := p(x^2)$ whenever $p \in \mathbb{R}[x]$, it follows that problem (2) is equivalent to solving

$$f^\star = \sup_{x\in\check{S}} \check{f}\,, \tag{4}$$

with

$$\check{S} := \{x \in \mathbb{R}^n \,:\, \check{g}_j(x) \ge 0\,,\, j \in [m]\}\,. \tag{5}$$

In the case of PMSV, one has $m = 2$, $\check{g}_1 = 1 - \sum_{i\in[n]} x_i^4$, $\check{g}_2 = 1$ and $\check{f} = (x^2)^\top M^\top M x^2$.

In Dickinson and Povh (2015), the authors state a specific constrained version of Pólya's Positivstellensatz. Explicitly, if $f, g_1,\dots,g_m$ are homogeneous polynomials, S is defined as in (3), and f is positive on $S\backslash\{0\}$, then

$$(\sum_{j\in[n]} x_j)^k f = \sum_{j\in[m]} \sigma_j g_j\,, \tag{6}$$

⋆ This work was supported by the Tremplin ERC Stg Grant ANR-18-ERC2-0004-01 (T-COPS project), the FMJH Program PGMO (EPICS project), as well as the PEPS2 Program (FastOPF project) funded by AMIES and RTE. This work has benefited from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Actions, grant agreement 813211 (POEMA) as well as from the AI Interdisciplinary Institute ANITI funding, through the French "Investing for the Future PIA3" program under the Grant agreement n°ANR-19-PI3A-0004.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

for some $k \in \mathbb{N}$ and homogeneous polynomials σ_j with positive coefficients. They also construct a hierarchy of linear relaxations associated with (6) converging from above to $f^\star$.

Contributions. Here, we extend this result to the case where the input polynomials f, g_j are all even, i.e., invariants under any variable sign flip, to get a representation similar to (6), where each multiplier σ_j is a sum of s-nomial squares. The case $s = 1$ corresponds to monomials, yielding a hierarchy of linear programming relaxations similar to Dickinson and Povh (2015), the case $s = 2$ corresponds to binomials, yielding a hierarchy of second-order conic programming relaxations. Fixing $s > 2$ in advance leads to a hierarchy of SDP relaxations with controlled sizes and allows us to overcome the potential ill-conditioning of the linear relaxations. Our optimization framework can be applied in particular to the PMSV problem and we illustrate the related numerical performance at the end of the paper.

Related works. A well-known method to approximate $f^\star$ is to consider the hierarchy of SDP relaxations by Lasserre (2001) based on the representation by Putinar (1993). Namely if f is positive on S and S involves the polynomial $g_1 = L - \|x\|_2^2$ for some $L > 0$, then $f = \sum_{j\in[m]} \sigma_j g_j$ for some SOS polynomials σ_j. The resuling hierarchy of SDP relaxations to approximate $f^\star$ from above is:

$$\inf \left\{ \lambda : \lambda - f = \sum_{j\in[m]} \sigma_j g_j \,, \deg(\sigma_j g_j) \le 2k \right\} . \tag{7}$$

Increasing k imrpoves the accuracy of the resulting upper bounds but the size of the SDP becomes rapidly intractable. To overcome this computational burden, a remedy consists of exploiting sparsity of the input polynomials; see the work by Waki et al. (2006) and Wang et al. (2021a, 2020). Another research direction is to restrict each multiplier σ_j to be a sum of s-nomials. When $s \in \{1, 2\}$, we retrieve the framework by Ahmadi and Majumdar (2019) based on scaled diagonally dominant sums of squares (SDSOS), generalized later on by Gouveia et al. (2022) for arbitrary larger s. However, the resulting hierarchy of linear or second-order conic programming relaxations does not produce a sequence of upper bounds converging to $f^\star$. It turns out that convergence can be obtained by multiplying f by the power of a given positive polynomial, e.g., $1 + \|x\|_2^2$, or equivalently forcing the multipliers σ_j to have denominators. Such representations have been derived in Pólya (1928); Reznick (1995); Putinar and Vasilescu (1999) and the latter led to the optimization framework presented in Mai et al. (2021).

2. EXTENDING PÓLYA'S REPRESENTATION

We now state our main result, which extends Pólya's representation to even input polynomials. In addition, we provide a degree bound for the multipliers.

Theorem 1. Let $g_1, \dots, g_m$ be even polynomials such that $g_1 = L - \|x\|_2^2$ for some $L > 0$ and $g_m = 1$. Let S be the semialgebraic set defined by

$$S := \{x \in \mathbb{R}^n \,:\, g_1(x) \ge 0 \dots, g_m(x) \ge 0\} . \tag{8}$$

Let f be an even polynomial and nonnegative on S and $d_f := \lfloor \deg(f)/2 \rfloor + 1$. Then the following statements hold:

(1) For all $\varepsilon > 0$, there exists $K_\varepsilon \in \mathbb{N}$ such that for all $k \ge K_\varepsilon$, there exist sums of monomial squares σ_j with $\deg(\sigma_j g_j) \le 2(k + d_f)$, $j \in [m]$ and

$$(1 + \|x\|_2^2)^k (f + \varepsilon) = \sum_{j\in[m]} \sigma_j g_j \,. \tag{9}$$

(2) If S has nonempty interior, there exist positive constants $\bar{c}$ and c depending on f and S such that for all $\varepsilon > 0$, (9) holds with $K_\epsilon = \bar{c}\varepsilon^{-c}$.

Theorem 1 can be proved in the same way as (Mai and Magron, 2022, Corollary 2). It is important to note that the theorem still holds if we replace the first constraint polynomial $g_1 = L - \|x\|_2^2$ by $L - \sum_{i\in[n]} x_i^4$ and we do so for the PMSV problem. If we remove the multiplier $(1+\|x\|_2^2)^k$ in (9), Theorem 1 is no longer true. Indeed, let $n = 1$, $f := (x^2 - \frac{3}{2})^2$ and assume that $f = \sigma_1(1-x^2)+\sigma_2$ for some SOS of monomials σ_1, σ_2. Note that f is even and positive on $[-1, 1]$. We write $\sigma_i := a_i + b_i x^2 + x^4 r_i(x)$ for some $a_i, b_i \in \mathbb{R}_+$ and $r_i \in \mathbb{R}[x]$. It implies that

$$\begin{aligned} x^4 - 3x^2 + \frac{9}{4} = {} & (a_1 + b_1 x^2 + x^4 r_1(x))(1 - x^2) \\ & + (a_2 + b_2 x^2 + x^4 r_2(x)) \,. \end{aligned} \tag{10}$$

Then we obtain the system of linear equations: $\frac{9}{4} = a_1 + a_2$ and $-3 = b_2 - a_1 + b_1$. Summing gives $-\frac{3}{4} = a_2 + b_2 + b_1$. However, $a_2 + b_2 + b_1 \ge 0$ since $a_i, b_i \in \mathbb{R}_+$, yielding a contradiction. Thus, Putinar's representation with sums of monomial squares does not exist for even input polynomials. With the multiplier $(1+x^2)^2$, we obtain a Pólya's representation:

$$(1 + x^2)^2 f = \bar{\sigma}_1 (1 - x^2) + \bar{\sigma}_2 \,, \tag{11}$$

where $\bar{\sigma}_1 = x^4 + \frac{15}{4} x^2 + \frac{9}{4}$ and $\bar{\sigma}_2 = x^8$ are SOS of monomials.

3. TRACTABLE SEMIDEFINITE RELAXATIONS

Given $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$, we write $x^\alpha = x_1^\alpha \cdots x_n^{\alpha_n}$. An s-nomial square is a polynomial which can be written as $(\sum_{i\in[s]} a_i x^{\alpha(i)})^2$, with $a_i \in \mathbb{R}$, $\alpha(i) \in \mathbb{N}^n$, $i \in [s]$. Let us denote by Σ_s the set of sums of s-nomial squares. The set of monomial squares corresponds to Σ_1. For any $s \in \mathbb{N}\backslash\{0\}$, one has the obvious inclusion $\Sigma_1 \subset \Sigma_s$, thus the representation result (9) from Theorem 1 holds with multipliers $\sigma_j \in \Sigma_s$, $j \in [m]$. With f and S as in Theorem 1, we rely on this representation to derive a converging sequence of upper bounds for $f^\star$. Each upper bound is obtained by solving an SDP, indexed by $s \in \mathbb{N}\backslash\{0\}$ and $k \in \mathbb{N}$:

$$\begin{aligned} \rho_{k,s} := \inf_\lambda \; & \lambda \\ \text{s.t. } & (1 + \|x\|_2^2)^k (\lambda - f) = \sum_{j\in[m]} \sigma_j g_j \,, \\ & \deg(\sigma_j g_j) \le 2(k + d_f) \,,\; \sigma_j \in \Sigma_s \,, j \in [m] \,. \end{aligned} \tag{12}$$

As a consequence of Theorem 1, we obtain the following convergence result.

Corollary 2. Let f and S be as in Theorem 1. The following statements hold:

(1) For all $k \in \mathbb{N}$ and for every $s \in \mathbb{N}\backslash\{0\}$, $\rho_{k,1} \geq \rho_{k,s} \geq f^\star$.
(2) For every $s \in \mathbb{N}\backslash\{0\}$, the sequence $(\rho_{k,s})_{k\in\mathbb{N}}$ converges to $f^\star$.
(3) If S has nonempty interior, there exist positive constants $\bar{c}$ and c depending on f and S such that for every $s \in \mathbb{N}\backslash\{0\}$ and for every $k \in \mathbb{N}$,

$$0 \leq \rho_{k,s} - f^\star \leq \left(\frac{k}{\bar{c}}\right)^{-\frac{1}{c}}. \tag{13}$$

Remark 3. We briefly outline the computational cost of SDP (12). In the case without constraints, one can show that the number of decision variables of is $\binom{n+k}{n} \cdot \binom{s}{2}$, to be compared with $\frac{1}{2}\binom{n+k}{n} \cdot \left(\binom{n+k}{n} + 1\right)$ in the dense case. Therefore, one expects a computational benefit when $s \ll \sqrt{\binom{n+k}{n}}$.

4. NUMERICAL EXPERIMENTS

We demonstrate the accuracy and efficiency of our optimization framework, namely the SDP relaxations (12) based on the extension of Pólya's representation theorem. Our framework is implemented in Julia 1.3.1 and available online via the link `https://github.com/maihoanganh/InterRelax`. We compare the upper bounds and runtime with the ones of the standard SDP relaxations (7), based on Putinar's Positivstellensatz. These latter relaxations are modeled by TSSOS Wang et al. (2021b). All SDP relaxations are solved by the solver Mosek 9.1. We use a desktop computer with an Intel(R) Core(TM) i7-8665U CPU @ 1.9GHz × 8 and 31.2 GB of RAM. Our benchmarks come from induced norm analysis of linear time-invariant systems, where the input signals are restricted to be nonnegative; we generate random matrices M as in (Ebihara et al., 2021, (12)), that is

$$M := \begin{bmatrix} D & 0 & 0 & \dots & 0 \\ CB & D & 0 & \dots & 0 \\ CAB & CB & D & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ CA^{r-2}B & CA^{r-3}B & CA^{r-4}B & \dots & D \end{bmatrix}, \tag{14}$$

where A, B, C, D are square matrices of size r, fixed in advance. In our numerical experiments, we choose $r \in \{4, 5, 6, 7\}$ and every entry of A, B, C, D is taken uniformly in $(-1, 1)$. Note that the number of variables of the resulting POP is $n = r^2$. We associate an "Id" to each related SDP relaxation, in which we compute a decomposition into sums of s-nomial squares. For Putinar's representation, we take a small relaxation order $k \in \{1, 2\}$ while for the extension Pólya's representation we let $k = 0$, leading to decompositions without denominators. We indicate the data of each SDP program, namely "nmat" and "msize" correspond to the number of matrix variables and their largest size, "nscal" and "naff" are the number of scalar variables and affine constraints. The symbol "−" means that the corresponding computation aborted due to lack of memory. The most accurate upper bounds are emphasized in bold. The total running time (including modeling and solving time) is indicated in seconds.

The numerical results are displayed in Table 1. The SDP relaxations associated to our extension of Pólya's representation provide more accurate upper bounds and they are computed more efficiently.

Table 1. Upper bounds of PMSV of M.

Id	n	Putinar			Pólya			
		k	upper bound	time	k	s	upper bound	time
1	16	1	47.48	0.02	0	17	**30.18**	0.7
2		2	**30.18**	16				
3	25	1	168.44	0.04	0	26	**91.28**	0.9
4		2	**91.28**	877				
5	36	1	4759.12	0.2	0	37	**2462.03**	1
6		2	–	–				
7	49	1	1777.53	0.5	0	50	**970.20**	2
8		2	–	–				

Id	Putinar				Pólya			
	nmat	msize	nscal	naff	nmat	msize	nscal	naff
1	1	17	38	153	1	17	138	153
2	17	153	154	4845				
3	1	26	27	351	1	26	327	351
4	26	351	352	23751				
5	1	37	38	703	1	37	668	703
6	37	703	704	91390				
7	1	50	51	1275	1	50	1227	1275
8	50	1275	1276	292825				

REFERENCES

Ahmadi, A.A. and Majumdar, A. (2019). DSOS and SDSOS optimization: more tractable alternatives to sum of squares and semidefinite optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(2), 193–230.

Dickinson, P.J. and Povh, J. (2015). On an extension of Pólya's Positivstellensatz. *Journal of global optimization*, 61(4), 615–625.

Ebihara, Y., Waki, H., Magron, V., Mai, N.H.A., Peaucelle, D., and Tarbouriech, S. (2021). l2 induced norm analysis of discrete-time LTI systems for nonnegative input signals and its application to stability analysis of recurrent neural networks. *European Journal of Control*.

Gouveia, J., Kovačec, A., and Saee, M. (2022). On sums of squares of k-nomials. *Journal of Pure and Applied Algebra*, 226(1), 106820.

Lasserre, J.B. (2001). Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3), 796–817.

Mai, N.H.A., Lasserre, J.B., and Magron, V. (2021). Positivity certificates and polynomial optimization on non-compact semialgebraic sets. *Mathematical Programming*, 1–43.

Mai, N.H.A. and Magron, V. (2022). On the complexity of Putinar-Vasilescu's Positivstellensatz. *Journal of Complexity*, 72, 101663.

Pólya, G. (1928). Über Positive Darstellung von Polynomen. *Vierteljschr. Naturforsch. Ges. Zürich*, 73, 141–145.

Putinar, M. (1993). Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42(3), 969–984.

Putinar, M. and Vasilescu, F.H. (1999). Positive polynomials on semi-algebraic sets. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 328(7), 585–589.

Reznick, B. (1995). Uniform denominators in Hilbert's seventeenth problem. *Mathematische Zeitschrift*, 220(1), 75–97.

Waki, H., Kim, S., Kojima, M., and Muramatsu, M. (2006). Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity. *SIAM Journal on Optimization*, 17(1), 218–242.

Wang, J., Magron, V., and Lasserre, J.B. (2021a). Chordal-TSSOS: a moment-SOS hierarchy that exploits term sparsity with chordal extension. *SIAM Journal on Optimization*. Accepted for publication.

Wang, J., Magron, V., Lasserre, J.B., and Mai, N.H.A. (2020). CS-TSSOS: Correlative and term sparsity for large-scale polynomial optimization. *arXiv preprint arXiv:2005.02828*.

Wang, J., Magron, V., and Lasserre, J.B. (2021b). TSSOS: A Moment-SOS hierarchy that exploits term sparsity. *SIAM Journal on Optimization*, 31(1), 30–58.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On Semi-uniform Input-to-state Stability and Polynomial Input-to-state Stability ⋆

Masashi Wakaiki *

* *Graduate School of System Informatics, Kobe University, Nada, Kobe, Hyogo 657-8501, Japan (e-mail: wakaiki@ruby.kobe-u.ac.jp).*

Abstract: We introduce the notions of semi-uniform input-to-state stability and polynomial input-to-state stability for infinite-dimensional systems. A characterization of semi-uniform input-to-state stability is developed based on attractivity properties as in the uniform case. We also provide sufficient conditions for linear systems to be polynomially input-to-state stable.

Keywords: Infinite-dimensional systems, Input-to-state stability, Polynomial stability

AMS subject classifications: 47D06, 93C25, 93D09

1. INTRODUCTION

The concept of input-to-state stability (ISS), introduced by Sontag (1989) for ordinary differential equations, unifies internal stability with respect to initial states and external stability with respect to disturbances. For infinite-dimensional systems, ISS has been also studied intensively in recent years; see, e.g., the survey article of Mironchenko and Prieur (2020) and the book of Karafyllis and Krstic (2019). In this talk, we introduce the notions of semi-uniform ISS and its subclass, polynomial ISS.

Semi-uniform stability of operator semigroups implies the uniform asymptotic behavior of semigroup orbits with respect to initial values from the unit ball of the domain of the generator endowed with the graph norm. In this sense, semi-uniform stability lies between exponential stability and strong stability. We refer the reader to the overview of Chill et al. (2020) for the recent developments of semi-uniform stability. The motivation of introducing semi-uniform ISS is to bridge the gap between uniform ISS and strong ISS as in the semigroup case.

In this talk, we present results recently obtained by Wakaiki (2022). First, we characterize semi-uniform ISS by using attractivity properties. This result is the semi-uniform version of the characterizations of uniform/strong ISS in Theorems 5 and 12 of Mironchenko and Wirth (2018). By this characterization, we show that semi-uniform ISS implies strong ISS for bilinear systems. Next, we provide sufficient conditions for linear systems to be polynomially ISS. Under the sufficient conditions, the range of an input operator is restricted depending on the polynomial decay rate of $\|T(t)A^{-1}\|$, where $(T(t))_{t\geq 0}$ is the polynomially stable semigroup governing the state evolution of the system without inputs and A is its generator.

2. DEFINITIONS

Let X and U be Banach spaces with norm $\|\cdot\|$ and $\|\cdot\|_U$, respectively. Let $\mathcal{U}$ be a normed vector space contained in the space $L^1_{\mathrm{loc}}(\mathbb{R}_+, U)$ of all locally integrable functions from $\mathbb{R}_+$ to U. We denote by $\|\cdot\|_{\mathcal{U}}$ the norm on $\mathcal{U}$. Assume that $u(\cdot+\tau) \in \mathcal{U}$ and $\|u\|_{\mathcal{U}} \geq \|u(\cdot+\tau)\|_{\mathcal{U}}$ for all $u \in \mathcal{U}$ and $\tau \geq 0$.

Consider a semi-linear system with state space X and input space U:

$$\Sigma(A,F) \quad \begin{cases} \dot{x}(t) = Ax(t) + F\big(x(t), u(t)\big), & t \geq 0 \\ x(0) = x_0, \end{cases}$$

where A is the generator of a C_0-semigroup $(T(t))_{t\geq 0}$ on X, $F : X \times U \to X$ is a nonlinear operator, $x_0 \in X$ is an initial state, and $u \in \mathcal{U}$ is an input.

Definition 1. Suppose that for every $\tau > 0$, $f \in C([0,\tau], X)$, and $g \in \mathcal{U}$, the map $t \mapsto F(f(t), g(t))$ is integrable on $[0,\tau]$. For $\tau > 0$, a function $x \in C([0,\tau], X)$ is called *a mild solution of* $\Sigma(A,F)$ *on* $[0,\tau]$ if x satisfies the integral equation

$$x(t) = T(t)x_0 + \int_0^t T(t-s)F\big(x(s), u(s)\big)ds$$

for all $t \in [0,\tau]$. Moreover, we say that $x \in C(\mathbb{R}_+, X)$ is *a mild solution of* $\Sigma(A,F)$ *on* $\mathbb{R}_+$ if $x|_{[0,\tau]}$ is a mild solution of $\Sigma(A,F)$ on $[0,\tau]$ for all $\tau > 0$.

Throughout this talk, we consider only forward complete systems defined as follows.

Definition 2. The semi-linear system $\Sigma(A,F)$ is *forward complete* if there exists a unique mild solution of $\Sigma(A,F)$ on $\mathbb{R}_+$ for all $x_0 \in X$ and $u \in \mathcal{U}$.

We denote by $\phi(t, x_0, u)$ the unique mild solution of the forward complete semi-linear system $\Sigma(A,F)$ with initial state $x_0 \in X$ and input $u \in \mathcal{U}$, i.e.,

$$\phi(t, x_0, u) := T(t)x_0 + \int_0^t T(t-s)F\big(\phi(s, x_0, u), u(s)\big)ds$$

for $t \geq 0$.

Let $\mathcal{K}_\infty$ and $\mathcal{KL}$ be the sets of the classic comparison functions from nonlinear systems theory. For the forward complete semi-linear system $\Sigma(A,F)$, we introduce the no-

⋆ This work was supported by JSPS KAKENHI Grant Number JP20K14362.

tion of semi-uniform input-to-state stability. Before doing so, we recall the definition of uniform global stability.

Definition 3. The semi-linear system $\Sigma(A,F)$ is called *uniformly globally stable (UGS)* if the following two conditions hold:

(1) $\Sigma(A,F)$ is forward complete;
(2) there exist $\gamma,\mu\in\mathcal{K}_\infty$ such that
$$\|\phi(t,x_0,u)\|\le\gamma(\|x_0\|)+\mu(\|u\|_{\mathcal{U}})$$
for all $x_0\in X$, $u\in\mathcal{U}$, and $t\ge 0$.

The graph norm $\|\cdot\|_A$ of a linear operator $A:D(A)\subset X\to X$ is defined by $\|x\|_A:=\|x\|+\|Ax\|$ for $x\in D(A)$.

Definition 4. The semi-linear system $\Sigma(A,F)$ is called *semi-uniformly input-to-state stable (semi-uniformly ISS)* if the following two conditions hold:

(1) $\Sigma(A,F)$ is UGS;
(2) there exist $\kappa\in\mathcal{KL}$ and $\mu\in\mathcal{K}_\infty$ such that
$$\|\phi(t,x_0,u)\|\le\kappa(\|x_0\|_A,t)+\mu(\|u\|_{\mathcal{U}})$$
for all $x_0\in D(A)$, $u\in\mathcal{U}$, and $t\ge 0$.

In particular, if there exists $\alpha>0$ such that for all $r>0$, $\kappa(r,t)=O(t^{-1/\alpha})$ as $t\to\infty$, then $\Sigma(A,F)$ is called *polynomially input-to-state stable (polynomially ISS) with parameter* α.

3. CHARACTERIZATION OF SEMI-UNIFORM ISS

We establish a characterization of semi-uniform ISS. The following two semi-uniform attractivity properties are used for the characterization.

Definition 5. The forward complete semi-linear system $\Sigma(A,F)$ has the *semi-uniform limit property* if there exists $\mu\in\mathcal{K}_\infty$ such that the following statement holds: For all $\varepsilon,r>0$, there is $\tau=\tau(\varepsilon,r)<\infty$ such that for all $x_0\in D(A)$,
$$\|x_0\|_A\le r\ \ \wedge\ \ u\in\mathcal{U}$$
$$\Rightarrow\quad\exists t\le\tau:\|\phi(t,x_0,u)\|\le\varepsilon+\mu(\|u\|_{\mathcal{U}}).$$

Definition 6. The forward complete semi-linear system $\Sigma(A,F)$ has the *semi-uniform asymptotic gain property* if there exists $\mu\in\mathcal{K}_\infty$ such that the following statement holds: For all $\varepsilon,r>0$, there is $\tau=\tau(\varepsilon,r)<\infty$ such that for all $x_0\in X$ with $\|x_0\|_A\le r$ and all $u\in\mathcal{U}$,
$$t\ge\tau\quad\Rightarrow\quad\|\phi(t,x_0,u)\|\le\varepsilon+\mu(\|u\|_{\mathcal{U}}).$$

Theorem 7. The following statements on the semi-linear system $\Sigma(A,F)$ are equivalent:

1. *$\Sigma(A,F)$ is semi-uniformly ISS.*
2. *$\Sigma(A,F)$ is UGS and has the semi-uniform limit property.*
3. *$\Sigma(A,F)$ is UGS and has the semi-uniform asymptotic gain property.*

For linear systems and bilinear systems, semi-uniform ISS implies strong ISS introduced in Definition 13 of Mironchenko and Wirth (2018); see also Nabiullin and Schwenninger (2018) for strong ISS.

Theorem 8. Assume that the operator F of $\Sigma(A,F)$ satisfies one of the following conditions:

1. *There exists $B\in\mathcal{L}(U,X)$ such that $F(\xi,v)=Bv$ for all $\xi\in X$ and $v\in U$.*
2. *For all $\xi,\zeta\in X$ and $v\in U$,*
$$F(\xi-\zeta,v)=F(\xi,v)-F(\zeta,v).$$

Then semi-uniform ISS implies strong ISS for $\Sigma(A,F)$.

4. POLYNOMIAL ISS OF LINEAR SYSTEMS

We recall the definition of polynomially stable semigroups.

Definition 9. A C_0-semigroup $(T(t))_{t\ge0}$ on a Banach space X generated by $A:D(A)\subset X\to X$ is called *polynomially stable with parameter* $\alpha>0$ if $(T(t))_{t\ge0}$ is uniformly bounded, if $\sigma(A)$ is contained in the open left half-plane, and if $\|T(t)A^{-1}\|=O\left(t^{-1/\alpha}\right)$ as $t\to\infty$.

Consider a linear system
$$\Sigma_{\rm lin}(A,B)\quad\begin{cases}\dot x(t)=Ax(t)+Bu(t),\quad t\ge0\\ x(0)=x_0,\end{cases}$$
where A is the generator of a polynomially stable semigroup on X and $B\in\mathcal{L}(U,X)$. We provide a simple sufficient condition for $\Sigma_{\rm lin}(A,B)$ to be polynomially ISS, by restricting the range of the input operator B.

Proposition 10. Let X and U be Banach spaces. Suppose that A generates a polynomially stable semigroup with parameter $\alpha>0$ on X. If $B\in\mathcal{L}(U,X)$ satisfies $\mathrm{ran}(B)\subset D((-A)^\beta)$ for some $\beta>\alpha$, then $\Sigma_{\rm lin}(A,B)$ is polynomially ISS with parameter α for $\mathcal{U}=L^\infty(\mathbb{R}_+,U)$.

We obtain a refined sufficient condition for polynomial ISS when linear systems are consisted of diagonalizable generators and finite-rank input operators; see Section 2.6 of Tucsnak and Weiss (2009) for diagonalizable operators.

Theorem 11. Let X be a Hilbert space and let U be a Banach space. Suppose that A is a diagonalizable operator generating a polynomially stable semigroup with parameter $\alpha>0$ on X. If $B\in\mathcal{L}(U,X)$ is a finite-rank operator and satisfies $\mathrm{ran}(B)\subset D((-A)^\alpha)$, then $\Sigma_{\rm lin}(A,B)$ is polynomially ISS with parameter α for $\mathcal{U}=L^\infty(\mathbb{R}_+,U)$.

REFERENCES

Chill, R., Seifert, D., and Tomilov, Y. (2020). Semi-uniform stability of operator semigroups and energy decay of damped waves. *Philos. Trans. Roy. Soc. A*, 378, 20190614.

Karafyllis, I. and Krstic, M. (2019). *Input-to-State Stability for PDEs.* Cham: Springer.

Mironchenko, A. and Prieur, C. (2020). Input-to-state stability of infinite-dimensional systems: recent results and open questions. *SIAM Review*, 62, 529–614.

Mironchenko, A. and Wirth, F. (2018). Characterizations of input-to-state stability for infinite-dimensional systems. *IEEE Trans. Automat. Control*, 63, 1692–1707.

Nabiullin, R. and Schwenninger, F.L. (2018). Strong input-to-state stability for infinite-dimensional linear systems. *Math. Control Signals Systems*, 30, Art. no. 4.

Sontag, E.D. (1989). Smooth stabilization implies coprime factorization. *IEEE Trans. Automat. Control*, 34, 435–443.

Tucsnak, M. and Weiss, G. (2009). *Observation and Control of Operator Semigroups.* Basel: Birkhäuser.

Wakaiki, M. (2022). Semi-uniform input-to-state stability of infinite-dimensional systems. doi: https://doi.org/10.1007/s00498-022-00326-1. To appear *in Math. Control Signals Systems.*

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A data-driven formulation for balanced truncation of bilinear dynamical systems

Ion Victor Gosea* **Igor Pontes Duff***
Serkan Gugercin** **Christopher Beattie****

* *Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstrasse 1, 39106, Magdeburg, Germany (e-mail: {gosea,pontes}@mpi-magdeburg.mpg.de).*
** *Department of Mathematics and Computational Modeling and Data Analytics Division, Academy of Integrated Science, Virginia Tech, Blacksburg, VA 24061, USA (e-mail: {gugercin,beattie}@vt.edu).*

Abstract: We describe here a non-intrusive data-driven time-domain formulation of balanced truncation (BT) for bilinear control systems. We build on the recent method of Gosea et al. (2021) that recasts the classic BT method for linear time invariant systems as a data-driven method requiring only evaluations of either transfer function values or impulse responses. We extend the domain of applicability of this non-intrusive data-driven method to bilinear systems, arguably the simplest nontrivial class of weakly nonlinear systems.

Keywords: data-driven modeling, balanced truncation, model reduction, bilinear dynamics

1. INTRODUCTION

Bilinear dynamical systems (BDS) are an important class of weakly nonlinear systems that appear naturally in many applications, see, e.g., Mohler (1970, 1973); Al-Baiyat et al. (1993); Saputra et al. (2019); Qian and Zhang (2014). They may be described through a state-space realization given as

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}\,u(t) + \mathbf{M}\mathbf{x}(t)u(t), \quad y(t) = \mathbf{c}^T\mathbf{x}(t), \quad (1)$$

where $\mathbf{x}(t) : \mathbb{R} \to \mathbb{R}^n$ is the system state, $u(t) : \mathbb{R} \to \mathbb{R}$ is a (scalar) control input, and system matrices are given by $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{n\times n}$ and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$. For simplicity, we describe only the case of single-input single-output (SISO) BDS's; the analysis can be extended to multi-input multi-output (MIMO) scenarios without difficulty.

In many applications of interest, the BDS represented in (1) has large state-space dimension and one seeks to reduce the model order to mitigate computational burdens arising in subsequent simulation and control. Many systems theoretic model reduction methods have been extended to reducing BDS; see, e.g., Antoulas et al. (2016a); Benner and Breiten (2012); Benner and Damm (2011b); Flagg and Gugercin (2015); Goyal (2018) and references therein. Balanced truncation (BT) in particular Mullis and Roberts (1976); Moore (1981) has long been one of the gold standards of model reduction. Classic BT is an intrusive projection-based method, requiring access to internal realizations of system dynamics. In recent work Gosea et al. (2021), we have developed a data-driven formulation of BT (called QuadBT), that only requires access system response sampling (either through transfer function evaluation or impulse response observation), but requires neither state-space realizations nor observations of the system state. The goal of this note is to extend this data-driven reformulation of BT to BDS.

The extension of BT to bilinear systems is well established, and generally requires the extraction of system Gramians through the solution of generalized ("bilinear") Lyapunov equations (in lieu of the classical Lyapunov equations used for LTI systems). The key to extending both BT, as well as our data-driven reformulation of it, QuadBT, to bilinear systems relies on the *Volterra series* representation of the BDS system response. Under mild assumptions, one may develop the solution of (1) as $\mathbf{x}(t) = \sum_{k=1}^{\infty} \mathbf{x}_k(t)$ (see Rugh (1981)), where

$$\begin{aligned} \dot{\mathbf{x}}_1(t) &= \mathbf{A}\mathbf{x}_1(t) + \mathbf{b}u(t), \text{ and} \\ \dot{\mathbf{x}}_k(t) &= \mathbf{A}\mathbf{x}_k(t) + \mathbf{M}\mathbf{x}_{k-1}(t)u(t), \text{ for } k > 1. \end{aligned} \quad (2)$$

The output is similarly expressed: $y(t) = \sum_{k=1}^{\infty} y_k(t)$, with $y_k(t) = \mathbf{c}^T\mathbf{x}_k(t)$, for $k \geq 1$, and so, the initial BDS may be recast as a cascade of coupled linear subsystems.

2. DATA-DRIVEN BILINEAR BT

The Volterra formalism of (2) may be resolved via variation of parameters and collected into an explicit series representation as in (3). This *Volterra series* comprises an infinite series of multivariate convolution integrals, and the kernel associated with the kth term, $h_k(t, \tau_1, \tau_2, \ldots, \tau_k)$, is called the *$k$th triangular Volterra kernel*.

Assigning $\tau_0 = t$ and introducing the change of variables $t_{k-i} = \tau_i - \tau_{i+1}$, for $i = 0, 1, \ldots, k-1$, one obtains the so-called *kth regular Volterra kernel*, which we write without a change in notation as

$$h_k(t_1, t_2, \ldots, t_k) = \mathbf{c}^T e^{\mathbf{A}t_k}\mathbf{M}e^{\mathbf{A}t_{k-1}} \cdots \mathbf{M}e^{\mathbf{A}t_1}\mathbf{b}, \quad k \geq 1.$$

A multivariate Laplace transform of $h_k(t_1, \ldots, t_k)$ yields the multivariate transfer function:

$$\begin{aligned} H(s_1, \ldots, s_k) = \mathbf{c}^T(s_k\mathbf{I} - \mathbf{A})^{-1}\mathbf{M}(s_{k-1}\mathbf{I} - \mathbf{A})^{-1}\mathbf{M}\cdots \\ \cdots \mathbf{M}(s_2\mathbf{I} - \mathbf{A})^{-1}\mathbf{M}(s_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}. \end{aligned}$$

$$y(t) = \sum_{k=1}^{\infty} \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_{k-1}} \underbrace{\mathbf{c}^T e^{\mathbf{A}(t-\tau_1)} \mathbf{M} e^{\mathbf{A}(\tau_1-\tau_2)} \cdots \mathbf{M} e^{\mathbf{A}(\tau_{k-1}-\tau_k)} \mathbf{b}}_{h_k(t,\tau_1,\tau_2,\ldots,\tau_k)} u(\tau_k) \cdots u(\tau_1)\, d\tau_k \cdots d\tau_1. \tag{3}$$

Not surprisingly, $h_k(t_1, \ldots, t_k)$ and $H(s_1, \ldots, s_k)$ extend the usual univariate impulse response and associated transfer function from linear problems to nonlinear problems and so form a basic building block for extending linear system-theoretic approaches for data-driven modeling to nonlinear dynamical systems.

In the LTI case (i.e., $\mathbf{M} = \mathbf{0}$ in (1)), the Gramians are defined in the time domain as $\mathbf{P} = \int_0^\infty e^{\mathbf{A}t}\mathbf{b}\mathbf{b}^T e^{\mathbf{A}^T t} dt$ and $\mathbf{Q} = \int_0^\infty e^{\mathbf{A}^T t}\mathbf{c}\mathbf{c}^T e^{\mathbf{A}t} dt$. For bilinear systems, the corresponding algebraic Gramians are based on the Volterra series expansion for $y(t)$ and are naturally defined as (Al-Baiyat and Bettayeb (1993); D'Alessandro et al. (1974))

$$\mathbf{P} = \sum_{k=1}^{\infty} \int_0^\infty \cdots \int_0^\infty \mathbf{z}_k \mathbf{z}_k^T dt_1 \ldots dt_k, \tag{4}$$

where

$$\begin{cases} \mathbf{z}_1(t_1) = e^{\mathbf{A}t_1}\mathbf{b}, & \text{for } k = 1, \\ \mathbf{z}_k(t_1, \ldots, t_k) = e^{\mathbf{A}t_k}\mathbf{M}\mathbf{z}_{k-1}, & \text{for } k \geq 2, \end{cases} \tag{5}$$

with analogous expressions for $\mathbf{Q}$. Under conditions sufficient to guarantee convergence of these infinite sums Zhang and Lam (2002), the Gramians $\mathbf{P}$ and $\mathbf{Q}$ solve the "bilinear Lyapunov equations"

$$\begin{aligned} \mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \mathbf{M}\mathbf{P}\mathbf{M}^T + \mathbf{b}\mathbf{b}^T = \mathbf{0} \qquad \text{and} \\ \mathbf{A}^T\mathbf{Q} + \mathbf{Q}\mathbf{A}^T + \mathbf{M}^T\mathbf{Q}\mathbf{M} + \mathbf{c}\mathbf{c}^T = \mathbf{0}. \end{aligned} \tag{6}$$

Assuming the system (1) is reachable and observable, then $\mathbf{P}$ and $\mathbf{Q}$ are positive definite matrices. See Zhang and Lam (2002) for details. The bilinear extension of the classical BT then proceeds as for the linear case. Let $\mathbf{U}, \mathbf{L} \in \mathbb{R}^{n\times n}$ be the square-root factors, i.e.,

$$\mathbf{P} = \mathbf{U}\mathbf{U}^T \quad \text{and} \quad \mathbf{Q} = \mathbf{L}\mathbf{L}^T. \tag{7}$$

Then, pick a truncation index, $1 \leq r \leq n$ and compute the SVD of the matrix $\mathbb{L} = \mathbf{L}^T\mathbf{U}$, which is then partitioned as

$$\mathbb{L} = \mathbf{L}^T\mathbf{U} = [\mathbf{Z}_1 \ \mathbf{Z}_2] \begin{bmatrix} \mathbf{S}_1 & \\ & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1^T \\ \mathbf{Y}_2^T \end{bmatrix}, \quad \begin{matrix} \mathbf{S}_1 \in \mathbb{R}^{r\times r}, \\ \mathbf{S}_2 \in \mathbb{R}^{(n-r)\times(n-r)}. \end{matrix} \tag{8}$$

The singular values of $\mathbb{L} = \mathbf{L}^T\mathbf{U}$ (i.e., the diagonal entries of $\mathsf{diag}(\mathbf{S}_1, \mathbf{S}_2)$ in (8)) are called the *Hankel singular values* of the underlying bilinear system. These values are system invariants, i.e., they are independent of realization. BT proceeds by truncating system states that correspond to small Hankel singular values in $\mathbf{S}_2$. Model reduction bases are then constructed from $\mathbf{W}_r = \mathbf{L}\mathbf{Z}_1\mathbf{S}_1^{-1/2}$ and $\mathbf{V}_r = \mathbf{U}\mathbf{Y}_1\mathbf{S}_1^{-1/2}$. As a consequence, the reduced order bilinear model can be constructed as

$$\begin{aligned} \hat{\mathbf{A}} &= \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r = \mathbf{S}_1^{-1/2}\mathbf{Z}_1^T(\mathbf{L}^T\mathbf{A}\mathbf{U})\mathbf{Y}_1\mathbf{S}_1^{-1/2}, \\ \hat{\mathbf{M}} &= \mathbf{W}_r^T\mathbf{M}\mathbf{V}_r = \mathbf{S}_1^{-1/2}\mathbf{Z}_1^T(\mathbf{L}^T\mathbf{M}\mathbf{U})\mathbf{Y}_1\mathbf{S}_1^{-1/2} \\ \hat{\mathbf{b}} &= \mathbf{W}_r^T\mathbf{b} = \mathbf{S}_1^{-1/2}\mathbf{Z}_1^T(\mathbf{L}^T\mathbf{b}), \\ \text{and} \quad \hat{\mathbf{c}} &= \mathbf{V}_r^T\mathbf{c} = \mathbf{S}_1^{-1/2}\mathbf{Y}_1^T(\mathbf{U}^T\mathbf{c}). \end{aligned} \tag{9}$$

This provides a high-fidelity, input-independent, reduced-order bilinear model. However, this approach is *projection-based* and hence *intrusive*, a drawback that we intend to improve upon. The key to our approach is the observation is that BT for bilinear systems does not require explicit access either to $\mathbf{U}$ or to $\mathbf{L}$; instead it needs (the SVD of) $\mathbb{L} = \mathbf{L}^T\mathbf{U}$ in (8), and the related expressions $\mathbf{L}^T\mathbf{A}\mathbf{U}$, $\mathbf{L}^T\mathbf{M}\mathbf{U}$, $\mathbf{L}^T\mathbf{b}$, and $\mathbf{U}^T\mathbf{c}$ in (9). We describe below how all these quantities can be approximated using only input/output data, without access to internal variables.

Even though the bilinear Lyapunov equations of (6) are linear in the unknowns, $\mathbf{P}$ and $\mathbf{Q}$, the terms $\mathbf{M}\mathbf{P}\mathbf{M}^T$ and $\mathbf{M}^T\mathbf{Q}\mathbf{M}$ create significant computational bottlenecks for large-scale problems. We refer the reader to, e.g., Kürschner (2016); Benner and Damm (2011a); Benner and Breiten (2013) and the references therein for some solution techniques.

One way of addressing this computational cost is to note that the terms in the infinite series (4) often decay rapidly and the whole series can be well approximated using only the leading two or three terms. This motivates the definition of truncated Gramians, $\mathbf{P}_T$ and $\mathbf{Q}_T$, which are obtained by truncating the series (4) after T terms. For example, if we pick a truncation index of $T = 2$, the corresponding truncated Gramian, $\mathbf{P}_T$, is

$$\begin{aligned} \mathbf{P}_T = &\int_0^\infty e^{\mathbf{A}t_1}\mathbf{b}\mathbf{b}^T e^{\mathbf{A}^T t_1} dt_1 \\ &+ \int_0^\infty \int_0^\infty e^{\mathbf{A}t_2}\mathbf{M}e^{\mathbf{A}t_1}\mathbf{b}\mathbf{b}^T e^{\mathbf{A}^T t_1}\mathbf{M}^T e^{\mathbf{A}^T t_2} dt_1 dt_2, \end{aligned}$$

with a similar expression for $\mathbf{Q}_T$. $\mathbf{P}_T$ and $\mathbf{Q}_T$ can be computed recursively in the following way: Notice that $\mathbf{P}_T$ can be written as $\mathbf{P}_T = \mathbf{P}_1 + \mathbf{P}_2$, where $\mathbf{P}_1$ and $\mathbf{P}_2$ solve

$$\begin{aligned} \mathbf{A}\mathbf{P}_1 + \mathbf{P}_1\mathbf{A}^T + \mathbf{b}\mathbf{b}^T = \mathbf{0} \quad &\Longrightarrow \quad \mathbf{P}_1 = \mathbf{U}_1\mathbf{U}_1^T \\ \mathbf{A}\mathbf{P}_2 + \mathbf{P}_2\mathbf{A}^T + \mathbf{M}\mathbf{U}_1(\mathbf{M}\mathbf{U}_1)^T = \mathbf{0} \quad &\Longrightarrow \quad \mathbf{P}_2 = \mathbf{U}_2\mathbf{U}_2^T, \end{aligned} \tag{10}$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ are the square-root factors of $\mathbf{P}_1$ and $\mathbf{P}_2$. Note $\mathbf{P}_1$ and $\mathbf{P}_2$ can be written in time domain as

$$\mathbf{P}_1 = \int_0^\infty e^{\mathbf{A}t}\mathbf{b}\mathbf{b}^T e^{\mathbf{A}^T t} dt \quad \text{and} \tag{11}$$

$$\mathbf{P}_2 = \int_0^\infty e^{\mathbf{A}t}(\mathbf{M}\mathbf{U}_1)(\mathbf{M}\mathbf{U}_1)^T e^{\mathbf{A}^T t} dt. \tag{12}$$

This provides a direct path forward for us, since we can use numerical quadratures on (11) and (12) to approximate the square-roots factors $\mathbf{U}_1$ and $\mathbf{U}_2$. Let $\widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{U}}_2$ be the approximate square roots obtained via numerical quadrature on (11) and (12), respectively. Then, an approximate square-root factor $\widetilde{\mathbf{U}}$ for $\mathbf{P}_T$ can be represented in terms of $\widetilde{\mathbf{U}}_1$ and $\widetilde{\mathbf{U}}_2$:

$$\begin{aligned} \mathbf{P}_T = \mathbf{P}_1 + \mathbf{P}_2 &\approx \widetilde{\mathbf{U}}_1\widetilde{\mathbf{U}}_1^T + \widetilde{\mathbf{U}}_2\widetilde{\mathbf{U}}_2^T \\ &= \begin{bmatrix} \widetilde{\mathbf{U}}_1 & \widetilde{\mathbf{U}}_2 \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{U}}_1 & \widetilde{\mathbf{U}}_2 \end{bmatrix}^T = \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T. \end{aligned} \tag{13}$$

One can similarly define a truncated observability Gramian $\mathbf{Q}_T$, which can be written as $\mathbf{Q}_T = \mathbf{Q}_1 + \mathbf{Q}_2$ (analogous to (10)) with the corresponding time-domain formula (analogous to (11) and (12)). Let $\widetilde{\mathbf{L}}_1$ and $\widetilde{\mathbf{L}}_2$ be the approximate square factors for $\mathbf{Q}_1$ and $\mathbf{Q}_2$, respectively, obtained via a numerical quadrature. Then

$$\begin{aligned}\mathbf{Q}_T = \mathbf{Q}_1 + \mathbf{Q}_2 &\approx \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1^T + \widetilde{\mathbf{L}}_2\widetilde{\mathbf{L}}_2^T \\ &= \begin{bmatrix}\widetilde{\mathbf{L}}_1 & \widetilde{\mathbf{L}}_2\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{L}}_1 & \widetilde{\mathbf{L}}_2\end{bmatrix}^T = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^T,\end{aligned} \tag{14}$$

thus given an approximate square-root factor $\widetilde{\mathbf{L}}$ for $\mathbf{Q}_T$.

Recall (8): BT requires forming the SVD of $\mathbb{L} = \mathbf{L}^T\mathbf{U}$. Approximate $\mathbb{L}$ using the approximate square-roots $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{L}}$; i.e.,

$$\widetilde{\mathbb{L}} = \widetilde{\mathbf{L}}^T\widetilde{\mathbf{U}} = \begin{bmatrix}\widetilde{\mathbf{L}}_1^T\widetilde{\mathbf{U}}_1 & \widetilde{\mathbf{L}}_1^T\widetilde{\mathbf{U}}_2 \\ \widetilde{\mathbf{L}}_2^T\widetilde{\mathbf{U}}_1 & \widetilde{\mathbf{L}}_2^T\widetilde{\mathbf{U}}_2\end{bmatrix}.$$

Our preliminary analysis below shows that as in data-driven BT for the LTI case Gosea et al. (2021), we will be able to evaluate $\widetilde{\mathbb{L}}$ (and thus its SVD) directly from input-output data.

To briefly illustrate this, apply a rectangular quadrature scheme to approximate the Gramians in (11) and (12) with two nodes in time, say τ_1 and τ_2. Then, we obtain,

$$\widetilde{\mathbf{U}_1} = [\mathbf{z}_1(\tau_1)\ \mathbf{z}_1(\tau_2)],$$
$$\widetilde{\mathbf{U}_2} = [\mathbf{z}_2(\tau_1,\tau_1)\ \mathbf{z}_2(\tau_1,\tau_2)\ \mathbf{z}_2(\tau_2,\tau_1)\ \mathbf{z}_2(\tau_2,\tau_2)],$$

where $\mathbf{z}_1(t_1) = e^{\mathbf{A}t_1}\mathbf{b}$ and $\mathbf{z}_2(t_1,t_2) = e^{\mathbf{A}t_1}\mathbf{M}e^{\mathbf{A}t_2}\mathbf{b}$, (with the goal of $\mathbf{P}_1 \approx \widetilde{\mathbf{U}}_1\widetilde{\mathbf{U}}_1^T$ and $\mathbf{P}_2 \approx \widetilde{\mathbf{U}}_2\widetilde{\mathbf{U}}_2^T$). Similarly,

$$\widetilde{\mathbf{L}_1} = [\mathbf{w}_1(\tau_1)\ \mathbf{w}_1(\tau_2)],$$
$$\widetilde{\mathbf{L}_2} = [\mathbf{w}_2(t_1,t_1)\ \mathbf{w}_2(\tau_1,\tau_2)\ \mathbf{w}_2(\tau_2,\tau_1)\ \mathbf{w}_2(\tau_2,\tau_2)],$$

where $\mathbf{w}_1(t_1) = e^{\mathbf{A}^Tt_1}\mathbf{c}^T$ and $\mathbf{w}_2(t_1,t_2) = e^{\mathbf{A}^Tt_1}\mathbf{M}^Te^{\mathbf{A}^Tt_2}\mathbf{c}^T$ (with the goal of $\mathbf{Q}_1 \approx \widetilde{\mathbf{L}}_1\widetilde{\mathbf{L}}_1^T$ and $\mathbf{Q}_2 \approx \widetilde{\mathbf{L}}_2\widetilde{\mathbf{L}}_2^T$). Hence, for this simple set-up, one can show that

$$\widetilde{\mathbf{L}_1}^T\widetilde{\mathbf{U}_1} = \begin{bmatrix}h_1(2\tau_1) & h_1(\tau_1+\tau_2) \\ h_1(\tau_1+\tau_2) & h_2(2\tau_2)\end{bmatrix},$$
$$\widetilde{\mathbf{L}_1}^T\widetilde{\mathbf{U}_2} = \begin{bmatrix}h_2(2\tau_1,\tau_1) & \dots & h_2(\tau_1+\tau_2,\tau_2) \\ h_2(\tau_1+\tau_2,\tau_1) & \dots & h_2(2\tau_2,\tau_2)\end{bmatrix},$$
$$\widetilde{\mathbf{L}_2}^T\widetilde{\mathbf{U}_1} = \begin{bmatrix}h_2(\tau_1,2\tau_1) & h_2(\tau_1,\tau_1+\tau_2) \\ \vdots & \vdots \\ h_2(\tau_2,\tau_1+\tau_2) & h_2(\tau_2,2\tau_2)\end{bmatrix},$$
$$\widetilde{\mathbf{L}_2}^T\widetilde{\mathbf{U}_2} = \begin{bmatrix}h_3(\tau_1,2\tau_1,\tau_1) & \dots & h_3(\tau_1,2\tau_2,\tau_2) \\ \vdots & \ddots & \vdots \\ h_3(\tau_2,\tau_2+\tau_1,\tau_2) & \dots & h_3(\tau_2,2\tau_2,\tau_2)\end{bmatrix}$$

As a consequence, the entries of $\widetilde{\mathbb{L}} = \widetilde{\mathbf{L}}^T\widetilde{\mathbf{U}}$ are solely determined from input/output data using only (time) samples of the subsystem kernels h_1, h_2 and h_3. In a similar way, one can also show that the other matrices in (9) appearing in BT, namely, $\widetilde{\mathbf{L}}^T\mathbf{A}\widetilde{\mathbf{U}}$, $\widetilde{\mathbf{L}}^T\mathbf{M}\widetilde{\mathbf{U}}$, $\widetilde{\mathbf{L}}^T\mathbf{b}$, and $\widetilde{\mathbf{U}}^T\mathbf{c}$ can be also be constructed relying only on kernels' data. Hence, by means of the SVD of $\widetilde{\mathbb{L}}$ as in (8), we are able to compute the matrices $\mathbf{Z}_1$ and $\mathbf{Y}_1$ and construct reduced models quantities in (9) directly from data. We call this framework QuadBT for bilinear balanced truncation. Due to the page limitations in this extended abstract, we have presented our formulation using only 2 quadrature nodes and unity weights. The general case follows immediately and will be included in the full paper.

We outlined our approach above using numerical quadrature in the time-domain but one can use an equivalent frequency-domain representation of the Gramians and consider instead quadrature rules in the frequency domain. Such a frequency-domain formulation would yield data matrices analogous to $\widetilde{\mathbb{L}}$, whose entries would now be derivable from the sampling of the multivariate subsystem transfer functions $H_k(s_1,\dots,s_k)$ defined above (and from associated divided differences).

The recent extensions of the Loewner framework applied to the class of BDS's in Antoulas et al. (2016b); Karachalios et al. (2021) make also use of specific transfer functions corresponding to the the BDS subsystems, i.e., regular transfer functions for the former and symmetric transfer functions for the latter. However, in these works, this is done in order to construct interpolatory models. We showed in Gosea et al. (2021) that our data-driven BT formulation outperforms the classical Loewner framework for LTI systems Mayo and Antoulas (2007) (in terms of the approximation errors). We anticipate a similar conclusion in the bilinear setting as well.

In our description given above, we have only considered the leading two terms of the Volterra series expansion. Although the leading two terms are enough in many cases Flagg and Gugercin (2015); Goyal (2018), we are investigating the impact of including higher-order Gramians, e.g., $\mathbf{P}_3, \mathbf{P}_4$ etc. We anticipate similar extensions that would require sampling of higher order kernels $h_k(t_1,\dots,t_k)$ and/or transfer functions $H_k(s_1,\dots,s_k)$.

3. PRELIMINARY NUMERICAL RESULTS

We consider the viscous Burgers' equation model from Benner and Breiten (2012). After applying a finite difference scheme for approximating the spatial derivatives, the resulting system of ODEs has a quadratic-bilinear nonlinearities with dimension n_q. By means of Carleman's linearization, a bilinear model of the form (1) with dimension $n = n_q^2 + n_q$ is obtained. To enforce positive-definite Gramians as the solutions of the Lyapunov equations (6), we multiply the matrix $\mathbf{M}$ with a positive scalar less than 1; here we choose $\gamma = 1/5 \in (0,1)$, i.e. $\tilde{\mathbf{M}} = \gamma\mathbf{M}$.

For a proof of concept example, we choose $n_q = 3$, leading to a bilinear system of dimension $n = 12$. Note that since the proposed framework is data-driven, the order of the underlying system does not play a role in our approach. We choose 50 linearly spaced time samples t_k in the interval $[0, 100]$s, truncation index $T = 2$, and construct the data-driven Loewner matrix $\widetilde{\mathbb{L}}$ whose singular values are expected to approximate the Hankel singular values. The results displayed in Fig. 1 verifies this expectation. The approximated Hankel singular values (obtained via the data-driven formulation) accurately match the original ones.

The Hankel singular value decay indicates that the underlying bilinear model is of minimal order $r_b = 7$. However, for MOR purposes, we use a slightly lower truncation order and choose $r = 5$. We simulate the full-model, the classical (projection-based) BT reduced model, and the data-driven QuadBT model with a control input $u(t) = 20(\cos(2\pi t) + \sin(12\pi t)e^{-0.4t})$ over $t \in [0, 5]$s. The resulting outputs and the output errors shown in Fig. 2 illustrate that the proposed data-driven QuadBT method closely mimics the performance of the classical BT.

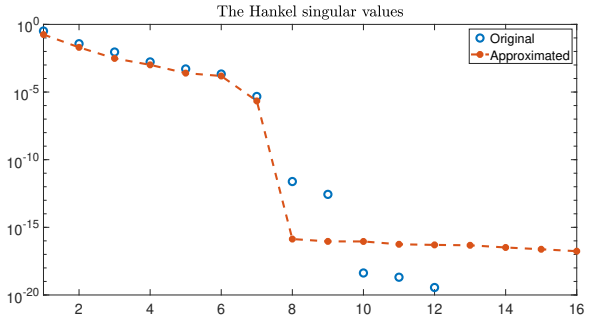


Fig. 1. Hankel singular values (original vs. approximated).

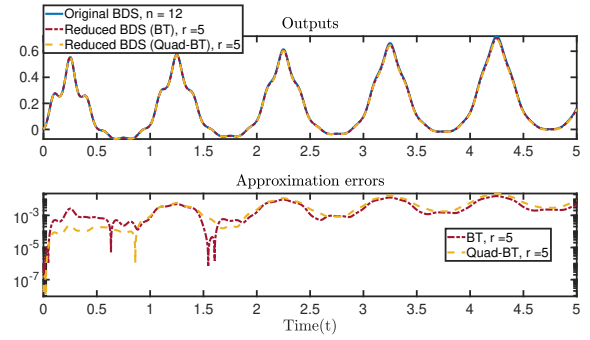


Fig. 2. Outputs (top) and output errors (bottom)

4. CONCLUSIONS

We developed a data-driven formulation for balanced truncation of bilinear systems, which does not require access to a state-space formulation and only works with the samples of subsystems kernels. The numerical results illustrate the potential of the new approach. Even though using a truncation index of $T = 2$ already yielded an accurate approximation, it would be interesting to include higher-order terms in the Volterra expansion, which will require sampling higher-order kernels. Since the Volterra kernels are smooth, we expect that progressively coarser sampling in each dimension will be sufficient for higher-order kernels and we anticipate significant benefits may accrue from sparse grid sampling and related methods.

REFERENCES

Al-Baiyat, S., Farag, A.S., and Bettayeb, M. (1993). Transient approximation of a bilinear two-area interconnected power system. *Electric Power Systems Research*, 26(1), 11–19. doi:10.1016/0378-7796(93)90064-L.

Al-Baiyat, S.A. and Bettayeb, M. (1993). A new model reduction scheme for k-power bilinear systems. In *Proceedings of 32nd IEEE conference on decision and control*, 22–27. IEEE.

Antoulas, A.C., Gosea, I.V., and Ionita, A.C. (2016a). Model reduction of bilinear systems in the Loewner framework. *SIAM J. Scientific Computing*, 38(5), B889–B916.

Antoulas, A.C., Gosea, I.V., and Ionita, A.C. (2016b). Model reduction of bilinear systems in the Loewner framework. *SIAM J. Scientific Computing*, 38(5), B889–B916.

Benner, P. and Breiten, T. (2012). Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. *SIAM J. Matrix Anal. Appl.*, 33, 859–885.

Benner, P. and Damm, T. (2011a). Lyapunov equations, energy functionals and model order reduction of bilinear and stochastic systems. *SIAM J. on Control and Optimization*, 49, 686–711.

Benner, P. and Breiten, T. (2013). Low rank methods for a class of generalized lyapunov equations and related issues. *Numerische Mathematik*, 124(3), 441–470.

Benner, P. and Damm, T. (2011b). Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. *SIAM journal on control and optimization*, 49(2), 686–711.

D'Alessandro, P., Isidori, A., and Ruberti, A. (1974). Realization and structure theory of bilinear dynamical systems. *SIAM Journal on Control*, 12(3), 517–535.

Flagg, G. and Gugercin, S. (2015). Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems. *SIAM J. Matrix Analysis and Applications*, 36, 549–579.

Gosea, I.V., Gugercin, S., and Beattie, C. (2021). Data-driven balancing of linear dynamical systems. *arXiv preprint arXiv:2104.01006*.

Goyal, P. (2018). *System-Theoretic Model Order Reduction for Bilinear and Quadratic-Bilinear Systems*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg.

Karachalios, D.S., Gosea, I.V., and Antoulas, A.C. (2021). On bilinear time-domain identification and reduction in the Loewner framework. In *Model Reduction of Complex Dynamical Systems*, volume 171 of *International Series of Numerical Mathematics*, 3–30. Birkhäuser, Cham. doi:10.1007/978-3-030-72983-7_1.

Kürschner, P. (2016). *Efficient low-rank solution of large-scale matrix equations*. Ph.D. thesis, Otto von Guericke Universität Magdeburg.

Mayo, A. and Antoulas, A. (2007). A framework for the solution of the generalized realization problem. *Linear Algebra and Its Applications*, 425(2-3), 634–662.

Mohler, R.R. (1970). Natural bilinear control processes. *IEEE Transactions on Systems Science and Cybernetics*, 6(3), 192–197. doi:10.1109/TSSC.1970.300341.

Mohler, R.R. (1973). *Bilinear Control Processes: With Applications to Engineering, Ecology and Medicine*, volume 106 of *Mathematics in Science and Engineering*. Academic Press, New York, London.

Moore, B. (1981). Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1), 17–32.

Mullis, C. and Roberts, R. (1976). Synthesis of minimum roundoff noise fixed point digital filters. *Circuits and Systems, IEEE Transactions on*, 23(9), 551–562.

Qian, K. and Zhang, Y. (2014). Bilinear model predictive control of plasma keyhole pipe welding process. *J. Manuf. Sci. Eng.*, 136(3), 031002. doi:10.1115/1.4025337.

Rugh, W.J. (1981). *Nonlinear System Theory: The Volterra/Wiener Approach*. Johns Hopkins University Press, Baltimore.

Saputra, J., Saragih, R., and Handayani, D. (2019). Robust H_∞ controller for bilinear system to minimize HIV concentration in blood plasma. *J. Phys.: Conf. Ser.*, 1245, 012055. doi:10.1088/1742-6596/1245/1/012055.

Zhang, L. and Lam, J. (2002). On $\mathcal{H}_2$ model reduction of bilinear systems. *Automatica*, 38(2), 205–216.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

3D Cues for Human Control of Target Acquisition in Auditory Augmented Reality ⋆

Konstantinos Dadamis, John Williamson, Roderick Murray-Smith

School of Computing Science, University of Glasgow, Scotland.
(e-mail: Roderick.Murray-Smith@glasgow.ac.uk).

Abstract: We compare the effectiveness of different auditory cues for attracting attention to spatial targets around a mobile user, using a commercial 3D audio headset instrumented with GPS and inertial sensors. We compare two approaches to spatial audio feedback with a baseline case that only provides 'on target' feedback: 1. hints as single sounds played from a 3D location and 2. frequency modulation of inter-pulse gaps based on proximity. We illustrate the difference in user control behaviour created by the different forms of feedback with phase plots. Single 3D sound hints provided the best improvement over the baseline case of no hint. Frequency modulation of pulses performed more poorly for larger targets. The choice of sound has a significant effect on targeting performance and there is a significant trade-off between efficient targeting and aesthetically-pleasing audio.

1. INTRODUCTION

Instrumented headsets which can sense orientation, location and bearing can be used to augment the user's experience of the world with a virtual audio layer. Fusing location awareness with orientation sensing allows accurate alignment of the virtual layer with real-world objects. In mobile contexts visual attention is a scarce resource, and navigation systems based on audio and vibrotactile cues Williamson et al. (2010); Holland et al. (2002) have successfully provided spatial guidance without overloading the visual channel. Positional audio could increase the efficiency of these navigation mechanisms.

Aside from the benefits of disengaging from the visual display, the advantages of positional audio cues are twofold. Firstly, audio cues extend the field of awareness of the user, presenting information close to their current location, but out of their current field of view, as discussed in Bolia et al. (1999). Secondly, audio cues function as effective attention management elements. Animation is an essential part of modern interfaces as it directs user attention to key UI components; the audio counterparts of animation cues can apply this attention management for entities out of view. We explore a range of possible solutions for effectively and efficiently informing the user about the location of nearby points of interest.

We used the Jabra Intelligent Headset [1] which includes 3D audio, GPS location, magnetometers, accelerometers and gyroscopes in iOS with the Jabra API. This commercially-available, integrated hardware package simplifies the equipment requirements for spatial audio target acquisition and provides a potential mass market for spatial audio applications.

2. TARGET ACQUISITION

There are several challenges when it comes to designing auditory cues for spatialised content. The cues need to be both efficient and result in a pleasant user experience. Key aspects of auditory target display are informing the user about the nature and number of targets nearby and their bearing and distance from the user. This paper explores different feedback mechanisms for informing users about the bearing of a single given target. The purpose of guidance feedback is to ease (quicker, requiring less effort) aligning head orientation with that of targets around the user. It needs to give a user hints about which direction to turn, and how close the target is. This feedback can be a single event (e.g. a "ping" in the target direction), while in others it is an ongoing process providing gradient information to ease acquisition. For a review of the spatial audio targeting literature see Marentakis (2006); Gröhn et al. (2005); Strachan et al. (2005); Sandberg et al. (2006); Eriksson (2008); McGookin et al. (2009). We explore three feedback conditions:

1. Baseline condition. No Hint In this experiment, the user is given no cue about the target direction, to explore the user's behaviour and performance on the simplest scenario, as a baseline. Instead, she only relies upon the simple feedback when on-target. Consequently, in order to acquire the target, she turns her head until feedback is heard.

2. Single-sound Hint from 3D location For each trial, a 3D pulse sound (with duration of 0.3-1.00s) is played once, from the direction of the target. We experiment with a number of different sounds.

3. Frequency modulation of pulses based on proximity In this case, the feedback given to the user simulated the behaviour of feedback from parking sensors available in many modern cars, where the delay between short pulses represents the distance to another car. Assume that the feedback pulse has a duration of τ, and the angular size of the target is w_t. Within the target area, feedback is played continuously, as in the previous experiments. When the user is in the opposite direction (180° from the target's centre), the pulse's period was set to be $k\tau$, where k is a specified multiplier. Consequently, when the user's distance from the target is $\psi \in (w_t/2, 180°]$, the pulse period is $\frac{\psi - w_t/2}{180° - w_t/2}(k-1)\tau + \tau$. k was set to 20 and the pulse used was a sinusoidal tone of 261Hz of duration $\tau = 0.1$s.

⋆ We acknowledge funding from GN Store Nord and EPSRC grant EP/R018634/1, *Closed-loop Data Science*

[1] **https://www.jabra.co.uk/supportpages/jabra-intelligent-headset**. Last accessed 25/1/2022.

As a pre-experiment, we investigated the sensitivity of faster cue-based acquisition to the specific sound used by exploring the impact of the types of sounds played on localisation speed. Our baseline was a simple sinusoidal tone (261Hz). When applying filters to white noise sounds, the widest filters (350-8000Hz) give the best localisation results Susnik et al. (2003), so the ideal sound should contain a wide range of frequencies. We tested two "blowing bottle" sounds from Cook (2002), and recorded two voiced vowel sounds. 4 further synthesized sounds were tested, Buzz-0004, Buzz-0035, Buzz-0036 and pulse1sec. 3D positional audio works best with sounds with strong transients and significant high-frequency content, for a clear inter-aural time delay and a perceptible effect of the head-related transfer function, which primarily modulates high-frequency components. As expected, the pure tone sound did not perform well, but surprisingly the richer "blowing bottle" sounds and the voiced sound performed worse. The sounds with fastest responses are Buzz-0004, Buzz-0036, the 'a' vowel recording and 'pulse1sec', which was the overall best. The aesthetic aspect of the sounds is important for the user experience, but the best performing sounds, apart from Buzz-0004, were considered to be somewhat robotic, squeaky or eccentric for a mainstream target acquisition application.

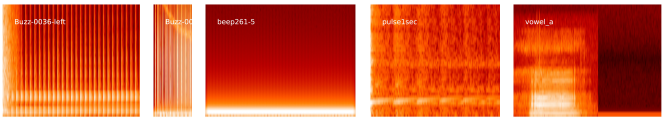


Fig. 1. Spectrograms of sounds used. Clipped to 8KHz max. freq, NFFT=1024, 1000 sample overlap.

3. EXPERIMENTAL SCENARIO

We investigate the impact of feedback choice on user behaviour, in terms of speed of action, nature of movement and user experience. In all cases, a simple 261Hz feedback tone indicates that the user is 'on-target'. The user is considered to be located at a fixed position, so that the GPS location uncertainty does not affect the results. As such the results provide a 'best case' scenario. The experimental task was performed on-campus, and the targets used are parts of the University of Glasgow's Main Building, as shown in Figure 2. Successful acquisition is defined as being achieved when the user looks towards the target bearing ψ_t for 3 seconds consecutively. The target is considered to be missed when the user has not acquired the target within 20 seconds. The targets are considered to be at the same distance d from the user and have the same angular w_t size (in degrees). We consider the user to be looking at the target when the direction is within the angle range $\psi_t \pm \frac{w_t}{2}$.

Fig. 2. Panoramic view of experiment location.

5 users aged between 23-46 and self-declared normal hearing tested the system. Each test for each of 7 target sizes w_t consists of the same ten targets in sequence, and the order of condition was cycled through participants. After acquiring or missing one target, the user was immediately presented with the next target. After all 10 targets at a given size w_t, the user rested for two minutes, then continued with the next sub-test of a different target size. Each experiment consists of 7 tests, for target sizes of w_t = 5, 10, 20, 40, 60, 90 and 180°, ordered from easiest (largest) to smallest, to give users progressively more challenging tasks. Within a particular feedback condition (e.g. frequency modulation of pulses) the order of the targets was kept constant. The users were not able to memorise the sequence. The sequence was varied across conditions. In most trials the users moved their whole body and not just the head in order to acquire the target. The sounds used are shown in the spectrogram plots in Figure 1.

4. ANALYSIS OF RESULTS

We view the user's behaviour from the perspective of a control system minimising the 'error' between the current bearing angle ψ and the reference or target bearing ψ_t, such that error $e = \psi_t - \psi$. The different feedback mechanisms will change the overall control system behaviour, Jagacinski and Flach (2003); Poulton (1974). In the 'no hint' case, feedback is only provided when over the target, so the user has an exploration behaviour. In the '3D hint' case feedback is provided once, at the start, to help the user infer target location so any error minimisation is being done by the user, with respect to the user's inferred target location. In the 'frequency modulation of pulses' method, explicit error feedback is provided in an ongoing fashion.

The experiments are sampled at 20Hz, and the data is smoothed using a Savistky-Golay filter (length 41 samples, order 4 polynomial), equivalent to least-squares polynomial fitting of a quartic polynomial to the last 2.05 seconds of data. This filter structure better preserves edges and transients than standard low-pass filters and can be used to robustly estimate derivatives.

The phase and polar plots shown in Fig. 3 provide a visual summary of acquisition performance. Target overshooting, oscillation and under-damped behaviour are all clearly visible. In contrast to time-series, phase plots make it easier to align and compare the dynamics of multiple acquisitions as time offsets are ignored. An example of an unsuccessful acquisition is illustrated in Figure 3a, where the user receives no hint about the location of the target, and fails to acquire it within the 20s time limit. The user enters and exits the target zone associated with the feedback tone starting and stopping, but continues to overshoot and 'hunt' around the small target.

4.1 No 3D hint

User performances are summarized in Figure 4, where the standard error of the mean time is indicated by error bars. The acquisition time decreases as the target size increases. No hint was slowest for all users apart from User 2 who's slowest condition was frequency modulation of pulses. The users missed significant numbers of targets on the 5 & 10° tests. Figure 3b shows a successful acquisition for a larger, easier target size of 20°. The user was initially close to the target (˜50°), but with no cue, chose the longer, slower way (310°).

4.2 With single sound 3D hint

Figure 4 shows a significant improvement in the acquisition times compared to the baseline case of *no hint*. The 3D hint that the user receives provides the approximate direction of the target and lets the user turn immediately towards that direction. As expected, the 3D audio made it clear whether a target is on the left or right of the user, but it was not as easy to perceive whether it was in front of or behind the user. The user's search for the target became faster on average for all users, with larger improvements for smaller targets.

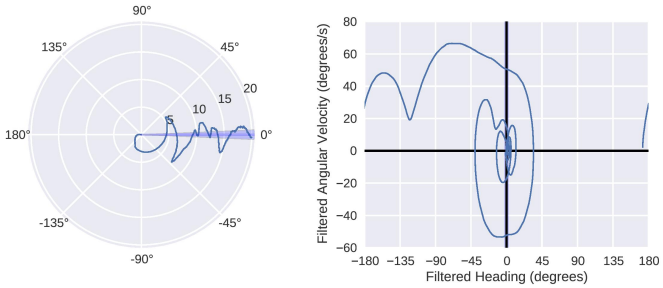
4.3 Frequency modulation of pulses based on proximity

Fig. 4 shows that for small targets (w_t 5-20°), this approach can speed responses over *No 3D hint*, but for larger targets (w_t 40-180°), it adds little, or gets worse. User 2 was slower throughout with this approach. Users had no initial hint of the target's direction, so to find the shortest path, some scanned the area around them, by quickly turning to the left and right (as shown in the edges of Fig. 5f), and then followed the path which increased the pulse frequency. Others went for one direction or the other until they heard the first target cues. The proximity indication of target reduces the velocity near the target.

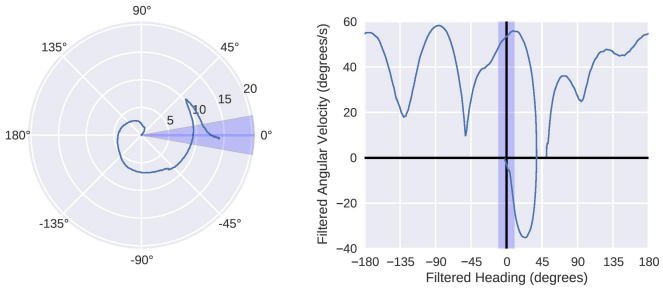
Summarising the statistics, comparing ratios of means for each type of trial, the relative speed up using 3D hints over no hint is 40% ($\mu = 1.40$, $\sigma = 0.55$). The speed up of 3D hints over the Frequency approach is 25% ($\mu = 1.25, \sigma = 0.74$). The No Hint case had most misses ($\mu = 1.39$), followed by Frequency ($\mu = 1.21$) and fewest was 3D hint ($\mu = 0.79$).

4.4 Comparison of Phase plots

The use of phase plots to represent the error convergence allows us to show multiple acquisitions of different time lengths on a single plot which allows us to test for consistent changes in approach depending on the feedback style. We have grouped the responses to small ($w_t = 5°$) targets and large ($w_t = 60°$) targets. Smaller targets in Figure 5 show underdamped responses where the user oscillates around the target, whereas larger targets show overdamped responses where the user hits the target and stays there. For larger initial errors, the velocity decrease slows already before the target zone is reached in Figure 5b, suggesting that the user enters a different control mode (akin to Costello's *Surge Model* Costello (1968)), however for larger targets, in Figure 5e there is less anticipatory change, and no further oscillatory control near the target, leading to larger final errors. The 3D hint led to smoother velocity profiles outside the target zone, suggesting that the user has a good sense of target location, where other feedback mechanisms, especially no feedback and pulse frequency modulation, have more variation in bearing velocity. Oscillation around the target is worst with no hint. Surges from initial conditions to close to the target are larger for the single sound hint (the 3 largest velocities when crossing 0° are for the 3 initial conditions closest to the target) suggesting scope for improving performance for nearby targets.



(a) Unsuccessful acquisition for target of size 5° without 3D hint



(b) Successful acquisition for target of size 20° without 3D hint

Fig. 3. User behaviour without an audio cue, during one target acquisition. Left: evolution of the bearing error angle e on polar axes. The radial part r represents time, the angle e is the user's bearing error. The $0 \pm \frac{w_t}{2}$ target zone is shaded to ease comparison. Right: phase plots of time derivative $\dot{e}$ against the user's bearing error e.

4.5 Movement time and difficulty of task

We investigated the relationship between the movement time and the index of difficulty for the "No 3D hint" and "With single sound 3D hint" systems. Meyer et al. Meyer et al. (1990), developing earlier work Crossman and Goodeve (1983), proposed that the time (MT) to move to a target area is a function of the distance to the target (A) and the size of the target (W), $MT = a + bID$, where the index of difficulty, $ID = (\frac{A}{W})^{\frac{1}{n}}$, where n relates to the upper limit on submovements. $n = 2.6$ minimised the RMS error. Figs. 6a and 6b show the linear relationship between the MT and ID. The circles' radii $r \propto W$. Blue circles indicate that the user did not go past 180°, while red circles indicate the user took the long way round, with higher MT. Large targets have lower ID and lower MT.

5. CONCLUSIONS

We demonstrate auditory targeting behaviour with a commercial, instrumented headset. The headset and API provided a practical development platform for spatial audio systems. Experimental results demonstrate that the use of 3D hints for auditory targeting is an improvement over no feedback. User feedback indicated that this provides a simple, intuitive, aesthetically pleasing way for users to locate targets and required the least mental and physical workload. The pulse-frequency approach was less effective, slowing users for larger targets.

Visualisation of experimental results based on phase plots standard in control applications, can aid the design of bearing-based interaction. Phase plots allow rapid comparison of behaviour from time-series of varying lengths and present a clear visual summary of the dynamics of target acquisition, where the distance cue leads to a ballistic 'surge' phase Costello (1968) where the user moves towards the target zone to the final control phase. The additional 'radar' visualisation approach to time-series representation gives a clear representation of the head movement during the acquisition process, highlighting areas of high activity, which are likely to lead to lower usability results.

REFERENCES

Bolia, R.S., D'Angelo, W.R., and McKinley, R.L. (1999). Aurally aided visual search in three-dimensional space. *Human Factors*, 41(4), 664–669.

Cook, P.R. (2002). *Real Sound Synthesis for Interactive Applications*. A K Peters, Wellesley, Massachusetts.

Costello, R. (1968). The surge model of the well-trained human operator in simple manual control. *IEEE Transactions on Man–Machine Systems*, 9(1).

Crossman, E.R.F.W. and Goodeve, P.J. (1983). Feedback control of hand-movement and fitts' law. *The Quarterly Journal of Experimental Psychology*, 35(2), 251–278.

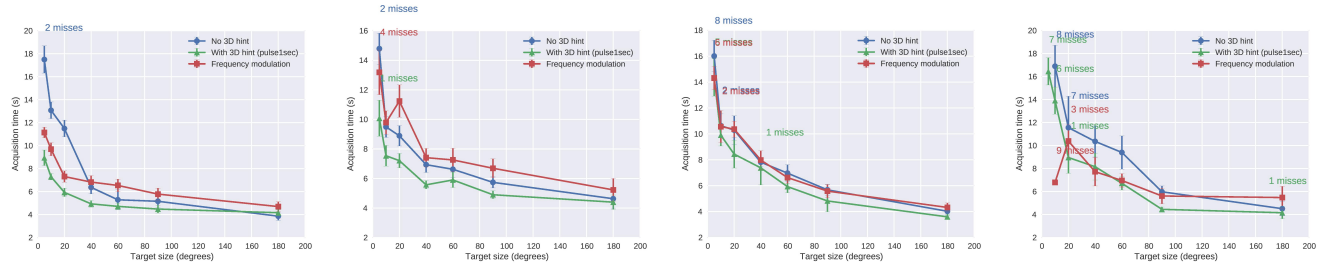
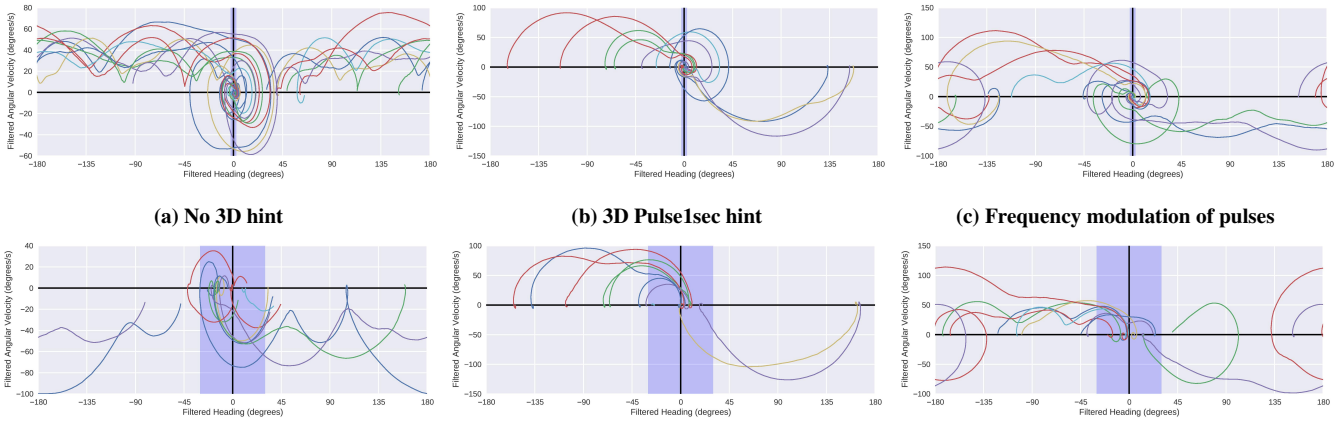


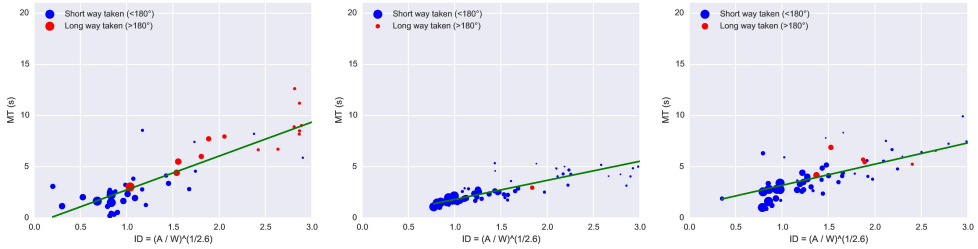
Fig. 4. Time-to-target results for users 1-4 in experiments with and without 3D hint and "Frequency modulation of pulses". Only successful selections are included in the mean & std. err. # of misses are shown beside each point.



(a) No 3D hint **(b) 3D Pulse1sec hint** **(c) Frequency modulation of pulses**

(d) No 3D hint **(e) 3D Pulse1sec hint** **(f) Frequency modulation of pulses**

Fig. 5. Phase plots for User 1 for small ($w_t = 5°$, upper) and large ($w_t = 60°$, lower) targets. Blue = feedback zone.



(a) Without 3D hint. $a = -0.57, b = 3.31$ **(b) With single 3D Pulse1sec hint.** $a = -0.06, b = 1.86$ **(c) Frequency modulation of pulses.** $a = 1.11, b = 2.07$

Fig. 6. Meyer's Power Law analysis of successful targeting for User 1. Movement time vs Index of Difficulty

Eriksson, L.*et al.*. (2008). On visual, vibrotactile, and 3D audio directional cues for dismounted soldier waypoint navigation. In *Proc. HF & Erg. Soc.*, volume 52, 1282–1286. SAGE.

Gröhn, M., Lokki, T., and Takala, T. (2005). Comparison of auditory, visual, and audiovisual navigation in a 3D space. *ACM Trans. Applied Perception (TAP)*, 2(4), 564–570.

Holland, S., Morse, D.R., and Gedenryd, H. (2002). AudioGPS: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous computing*, 6(4), 253–259.

Jagacinski, R.J. and Flach, J.M. (2003). *Control Theory for Humans: Quantitative approaches to modeling performance*. Lawrence Erlbaum, Mahwah, New Jersey.

Marentakis, G. (2006). *Deictic Spatial Audio Target Acquisition in the Frontal Horizontal Plane*. Ph.D. thesis, Univ. Glasgow.

McGookin, D., Brewster, S., and Priego, P. (2009). Audio bubbles: Employing non-speech audio to support tourist wayfinding. In *HAID*, 41–50. Springer.

Meyer, D., Keith-Smith, J.E., Kornblum, S., Abrams, R.A., and Wright, C.E. (1990). Speed-accuracy trade-offs in aimed movements: Toward a theory of rapid voluntary action. *M. Jeannerod (Ed.), Attention and Performance XIII*, 173–226.

Poulton, E.C. (1974). *Tracking skill and manual control*. Academic Press, New York.

Sandberg, S., Håkansson, C., Elmqvist, N., Tsigas, P., and Chen, F. (2006). Using 3D audio guidance to locate indoor static objects. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, 1581–1584.

Strachan, S., Eslambolchilar, P., Murray-Smith, R., Hughes, S., and O'Modhrain, S. (2005). gpsTunes: controlling navigation via audio feedback. In *MobileHCI*, 275–278. ACM.

Susnik, R., Sodnik, J., and Tomazic, S. (2003). Sound source choice in HRTF acoustic imaging. *HCI Int. adj. proc.*, 101–2.

Williamson, J., Robinson, S., Stewart, C., Murray-Smith, R., Jones, M., and Brewster, S. (2010). Social gravity: a virtual elastic tether for casual, privacy-preserving pedestrian rendezvous. In *Proc. ACM SIGCHI*, 1485–1494.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The Role of Systems Theory in Control Oriented Learning

Mario Sznaier * **Alex Olshevsky** ** **Eduardo D. Sontag** *

* *ECE Department, Northeastern University, Boston, MA 02115.*
Emails: msznaier@coe.neu.edu, e.sontag@northeastern.edu
** *ECE, Boston University, Boston, MA 02215. Email: alexols@bu.edu*

Abstract: This extended abstract discusses the role that systems theory plays in unveiling fundamental limitations of learning algorithms and architectures when used to control a dynamical system, and in suggesting strategies for overcoming these limitations. As an example, a feedforward neural network cannot stabilize a double integrator using output feedback. Similarly, a recurrent NN with differentiable activation functions that stabilizes a non-strongly stabilizable system must be itself open loop unstable, a fact that has profound implications for training with noisy, finite data. A potential solution to this problem, motivated by results on stabilization with periodic control, is the use of neural nets with periodic resets, showing that indeed systems theoretic analysis is instrumental in developing architectures capable of controlling certain classes of unstable systems. The abstract finishes by arguing that when the goal is to learn control oriented models, the loss function should reflect closed loop, rather than open loop model performance, a fact that can be accomplished by using gap-metric motivated loss functions.

Keywords: Control Oriented Learning, Neural Nets, Reinforcement Learning.

1. INTRODUCTION AND MOTIVATION.

Despite recent advances in Machine Learning (ML), the goal of designing control systems capable of fully exploiting the potential of these methods remains elusive. Modern ML methods can leverage large amounts of data to learn powerful predictive models, but such models are not designed to operate in a closed-loop environment. Recent results on reinforcement learning offer a tantalizing view of the potential of a rapprochement between control and learning, but so far proofs of performance and safety are mostly restricted to limited cases (e.g. finite horizon LQR/LQG or iterative tasks). Thus, learning elements are often used as black boxes within the loop, with limited interpretability and less than completely understood properties. Further progress hinges on the development of a principled understanding of the limitations of control-oriented machine learning. This extended abstract presents some initial results unveiling the fundamental limitations of some popular learning algorithms and architectures when used to control a dynamical system. For instance, it shows that even though feed forward neural nets are universal approximators, they are unable to stabilize some simple systems. Along these lines we also show that a recurrent neural net with differentiable activation functions that stabilizes a non-strongly stabilizable system must be itself open loop unstable, and discuss the implications of this fact for training with noisy, finite data. On the other hand, this difficulty can be overcome by using either time varying architectures or architectures with periodic resets. We also present some empirical evidence that conventional, off the shelf Reinforcement Learning will fail to stabilize non-strongly stabilizable plants. The extended abstract finishes by arguing that when the goal is to learn stabilizing controllers, the loss function should reflect closed loop performance, a fact that can be accomplished by using gap-metric motivated loss functions.

[1] This work was partially supported by NSF grant CNS–2038493, AFOSR grant FA9550-19-1-0005, and ONR grant N00014-21-1-2431.

1.1 Fundamental limitations of Feed Forward NN.

Even though Feed Forward NN (FFNN) are routinely used as controllers, there are fundamental obstructions that may prevent the existence of stabilizing FFNN controllers with continuous activation functions (Sontag and Sussmann (1980)). In this portion of the extended abstract we present some simple examples illustrating these limitations.

Single Hidden Layer FNN. Recall the single hidden layer FNN can approximate arbitrarily well any continuous functions (Cybenko (1989). However, as shown in (Sontag (1992)), there exists an asymptotically controllable system that has the origin as a locally asymptotically stable equilibrium point of the zero input dynamics and yet it cannot be be stabilized on compact sets using a single hidden layer FNN, even with discontinuous activation functions. This limitation arises from the fact that the (one sided) inverse needed to implement a stabilizing controller cannot be generically approximated by a linear combinations of scalar functions of linear combinations, even when the forward mapping is continuous.

Similarly, single hidden layer FNN cannot control nonholonomic systems due to their inability to implement Lie Brackets. On the other hand, since continuous-time

periodic controllers can overcome topological obstructions (Khaneja and Brockett (1980)) we conjecture that if $\dot{x} = f(x) + g(x)u$ is stabilizable, there is a recurrent NN (RNN) with (continuous) activation ReLU, state z, and input x, and a feedback $u = k(z, x)$ so that $\{(0, z)\}$ is asymptotically stable.

The discussion above illustrates the limitations of single hidden layer FFNNs when used as controllers. However, this leaves open the question of whether multi-layer FFNN can be used as universal controllers. In the next section we show that this is not the case.

1.2 Inadequacy of Deep FFNNs for output feedback

In this section we illustrate with a simple example the limitations of FFNNs when used to implement output feedback controllers. To this effect, consider the stabilization of a double integrator using output feedback,

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= f(x_1)\,, \end{aligned} \tag{1}$$

where $f(x_1)$ is implemented by a FFNN. Such a controller can never render the origin a globally asymptotically stable equilibrium point. To see this, consider the "energy" function $V(x_1, x_2) = \frac{1}{2}x_2^2 - \int_0^{x_1} f(\lambda)d\lambda$. Since $dV/dt \equiv 0$, trajectories starting at any $(0, x_2)$,$x_2 \neq 0$ cannot asymptotically approach the origin $(0, 0)$. To be precise, suppose $f(x_1)$ is locally bounded and Lebesgue measurable. Then $F(\xi) := \int_0^{\xi} f(\lambda)d\lambda$ is locally Lipschitz, and $x_1(t)$ is absolutely continuous (a.c.), so also $F(x_1(t))$ is a.c., so V is a.c. Thus, the chain rule can be applied, and V is constant along trajectories. (The a.c. property rules out examples such as the Cantor function, where derivatives can be identically zero yet the function is not constant.)

1.3 RNNs and non-strongly stabilizable systems

This portion of the extended abstract discusses the challenges in using RNNs to control non-strongly stabilizable plants, that is, Linear Time Invariant (LTI) plants that cannot be stabilized by open loop stable LTI controllers. These plants are interesting both on their own and because their relationship to the problem of simultaneous stabilization (Doyle et al. (1992)). Recall that a SISO plant is strongly stabilizable if it satisfies the parity interlacing property Doyle et al. (1992): the number of real poles in the right half plane (RHP) (counted according to their multiplicity) in between every pair of RHP zeros (including those at infinity) is even.

Proposition 1. *If a RNN with differentiable activation functions stabilizes a non-strongly stabilizable plant, then the RNN must be open loop unstable.*

The proof follows by considering the controller obtained by linearizing the input/output (between time series) mapping implemented by the NN.

A more interesting case arises if we allow for recurrent NN that implement non-smooth mappings. To investigate this case, consider an ideal setting where a known, open loop unstable internally stabilizing controller is used to train the neural net (Fig. 1). This scenario arises for instance when seeking to optimize performance. In this situation,

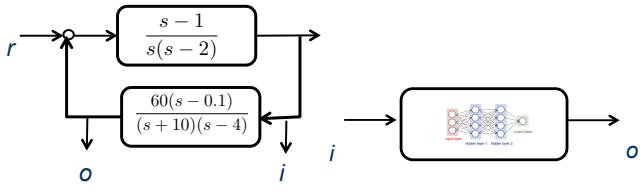


Fig. 1. Using the closed-loop signals generated by a stabilizing controller to train a NN.

one can use this pre-trained NN as an initial controller and then adjust its weights (for instance via gradient descent) to improve performance.

Let $i(t)$ and $o(t)$ denote the input output signals. Since the controller stabilizes the loop, it follows that $i(t), o(t) \in \mathcal{L}_\infty$, so in principle these bounded signals could be used to train a NN. Nevertheless, as we show next, a NN that interpolates all input/ouput pairs generated by an open loop unstable controller, has to be open loop unstable.

Proposition 2. *Consider an unbounded mapping $\mathcal{C}\colon \mathcal{L}_\infty \to \mathcal{L}_{\infty,e}$. Let $\mathcal{I}_b \subseteq \mathcal{L}_\infty$ denote the set of essentially bounded inputs that result in bounded outputs, i.e.*

$$\mathcal{I}_b \doteq \{r \in \mathcal{L}_\infty \colon (s \doteq \mathcal{C}r) \in \mathcal{L}_\infty\}$$

Then, if an operator $\mathcal{NN}$ is such that $\mathcal{NN}r = \mathcal{C}r$ a.e. for all $r \in \mathcal{I}_b$, $\mathcal{NN}$ must be open loop unstable, in the sense that there exists some $r_o \in \mathcal{L}_\infty$ such that $\mathcal{NN}r_o \notin \mathcal{L}_\infty$

From the observation above it follows that the NN can be trained in open-loop using the closed loop signals generated by an open loop controller only in the ideal case that these signals are perfectly known. This is a consequence of the fact that, since the NN is open-loop unstable, a suitably chosen perturbation of the input signal will lead to unbounded outputs. The discussion above leaves open the question of whether the NN can be trained in closed loop. As we show next, if the NN has differentiable activations, closed loop training is also likely to fail, due to the sensitivity of the parameters with respect to the observed outputs. For simplicity, we consider a SISO tracking scenario where the NN implements an LTI controller and the goal is to find the controller parameters θ that minimize some function $\mathcal{L}[e(\theta, u)]$ of the output e corresponding to a given input u, that is:

$$\theta^o = \operatorname*{argmin}_{\theta} \mathcal{L}[e(\theta, u)]$$

where

$$e(\theta) = \frac{u(s)}{1 + P(s)C(s, \theta)}$$

To illustrate the difficulties arising when the controller is open loop unstable, we will compute the gradient of the loss function with respect to the parameters of the controller. Let $C = \frac{N}{D}$ denote a coprime factorization of C, parameterized directly in term of its poles and zeros, that is $D(s) = D_o\Pi(s-\theta_i)$ and $N(s) = N_o\Pi(s-\psi_i)$. Since C is open loop unstable, at least one $\theta_i > 0$ and

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta_i} &= \frac{\partial \mathcal{L}}{\partial e}\frac{\partial e}{\partial C}\frac{\partial C}{\partial \theta_i} = \frac{\partial \mathcal{L}}{\partial e}\frac{-P}{(1+PC)^2}\frac{\partial C}{\partial \theta_i}u(s) \\ &= \frac{\partial \mathcal{L}}{\partial e}\frac{P}{(1+PC)^2}\left(\frac{C}{s-\theta_i}\right)u(s) \end{aligned} \tag{2}$$

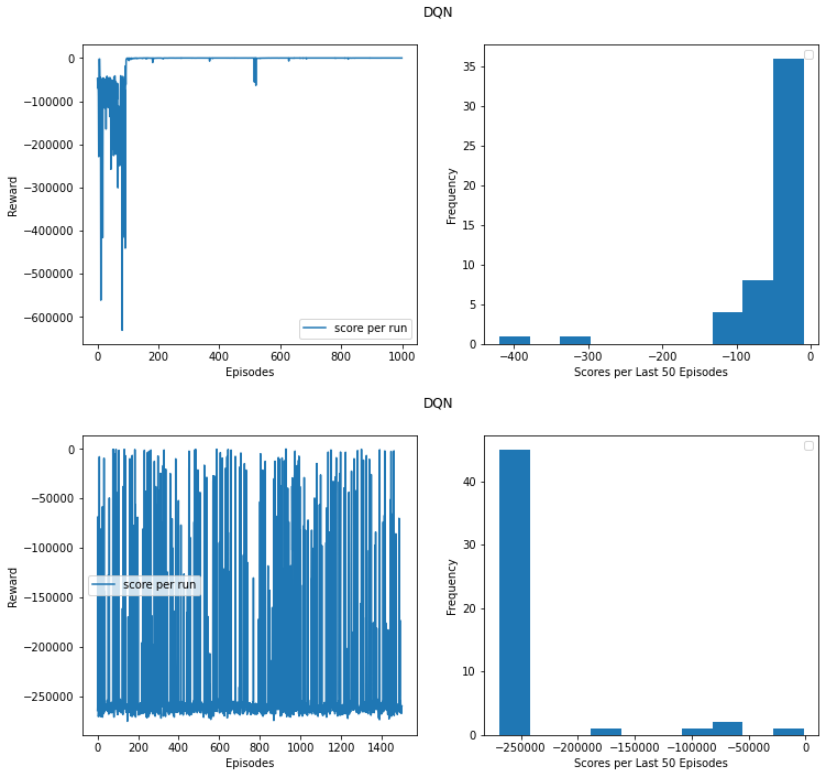


Fig. 2. Deep Reinforcement applied to a double integrator (top) and a non-strongly stabilizable plant (botton).

In the ideal case where $\frac{\partial C}{\partial \theta_i}$ can be exactly computed the poles of $\frac{\partial C}{\partial \theta_i}$ are cancelled by the zeros of $\frac{-P}{(1+PC)^2}$ and the overall system is stable. On the other hand, if only approximate values of the gradient are available (due for instance to finite and/or discrete time approximations), then this exact pole-zero cancellation no longer holds, leading to an unstable mapping $\frac{\partial C}{\partial \theta_i} \to \frac{\partial \mathcal{L}}{\partial \theta_i}$.

The developments above raise the question of whether a non-strongly stabilizable plant can be stabilized by an open loop stable controller. An affirmative answer to this question was given in Savkin and Petersen (1997), showing that this is indeed possible when using linear time varying, infinite dimensional controllers. An alternative, simpler controller is presented below:

Proposition 3. *Consider a non-strongly stabilizable LTI plant P and an LTI stabilizing controller with state space realization: A_c, B_c, C_c, D_c. Then the controller*

$$\mathcal{C}(y) = \begin{cases} \dot{x}_c = A_c x_c + B_c y \\ x_c(t^+) = x_c(t^-),\ t \neq kT \\ x_c(kT) = 0 \\ u = C_c x_c + D_c y \end{cases} \tag{3}$$

is open loop stable and stabilizes P.

Intuitively, the states of the controller are reset every T seconds to prevent them from growing too large. At the same time, since for $t \in (kT, (k+1)T)$ the LTI controller is acting, T can be chosen so that at the end of each cycle the state of the plant satisfies $\|x(kT+T)\| < \|x(kT)\|$. While in principle this avoids the difficulties entailed in training an open-loop unstable controller, at the moment is unclear how to implement and train such a controller using available NN architectures.

1.4 Reinforcement Learning

Next, we present some experiments illustrating the difficulties of using Reinforcement Learning to control non-strongly stabilizable plants. Consider the problem of stabilizing a plant using Deep Reinforcement Learning. To this effect, we considered a neural network architecture consisting of two hidden layers with leaky ReLu activations and a set of discrete actions $\mathcal{U}$. The NN takes an observation (i.e., $y_k = Cx_k$) and outputs a vector $q_k = V_\theta(y_k)$ of the same dimension as the number of actions, where each entry is a prediction of the value from taking the corresponding action. The next control action u_k is selected as the one corresponding to the maximal entry in q, with probability $1 - \epsilon_k$, or a random action with probability ϵ_k, where $\epsilon_k = \max\{\epsilon_{\min}, 0.99 * \epsilon_{k-1}\}$ The reward corresponding to the action u_k at state x_k is set to be $-\|x_k\|_2^2$.

The neural net was trained with Q-learning as follows. Let $u_{\text{taken,k}}$ denote the action taken at step k, and let q_{k+1} be the vector obtained by applying the neural network to the next observation $y_{k+1} = Cx_{k+1}$. We then set $\hat{q}_k$ to be the vector obtained by replacing the entry in q_k corresponding to $u_{\text{taken,k}}$ with $-||x_k||^2 + \gamma \max\{q_{k+1}\}$, where max applied to a vector denotes the largest entry. Finally, we perform a gradient descent step on θ, the weights of the NN, with objective $\|V_\theta(y_k) - \hat{q}_k\|^2$. Note that while knowledge of the true states was used in training (through the reward), the policy here depends only on the observations y_k. We applied this approach on both an "easy" plant (a double integrator with state feedback)

$$\begin{aligned} x(k+1) &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(k) \\ y(k) &= x(k) \end{aligned} \tag{4}$$

and a "hard" one (not strongly stabilizable)

$$x(k+1) = \begin{bmatrix} 1.2 & 0 \\ 0.1 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 0.1 \\ 0 \end{bmatrix} u(k) \qquad (5)$$
$$y(k) = [1 \ -1]\, x(k)$$

As shown in Fig. 2 the Deep RL algorithm described above stabilizes the "easy" case but fails to do so for the non-strongly stabilizable one.

1.5 Open Loop vs Closed Loop Distances

In this portion of the extended abstract we argue that when using a NN to model a plant, the loss function used to train it should take into account the closed-loop distance between the the unknown plant and its model, rather than the open loop one. Consider the open-loop unstable plant $G_1 = \frac{100}{2s-1}$. Modelling this plant with a NN such that the open loop distance, measured in terms of the induced norm $\|(G_1 - NN)\|_{\ell_i \to \ell_o}$ is finite, will require an open loop unstable net. On the other hand, when the loop is closed with the simple controller $K = 1$, the original plant G_1 and the open-loop stable plant $G_2 = \frac{100}{2s+1}$ have virtually indistinguishable performance (Fig 3) Thus, if the goal

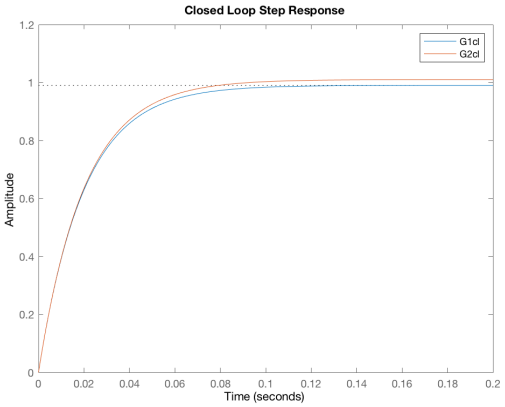


Fig. 3. Closed loop step responses of G_1 and G_2.

is to designing controllers, the stable plant G_2, which is substantially easier to model using a NN, can be used as a proxy for G_1 in the design process. This observation suggests that, when training a NN, one should try to minimize a closed-loop distance, rather than an open loop one. One such metric is the gap metric (see for instance Zhou and Doyle (1998)). Given two plants G_1, G_2 with normalized coprime factorizations $G_i = \frac{N_i}{D_i}$, $i = 1, 2$ the ν-gap δ_ν is defined by

$$\delta_\nu(G_1, G_2) = sup_w | - N_2(jw)M_1(jw) + \\ + M_2(jw)N_1(jw)|$$

Plants with small δ_ν can be stabilized by the same $\mathcal{H}_\infty$ optimal controller and have similar closed loop transfer functions (see Zhou and Doyle (1998) for a formalization of this statement). For instance, for the example above $\delta_\nu(G_1, G_2) = 0.02$, which explains the virtually indistinguishable closed loop responses. This suggest that one should learn coprime factorizations, rather than plants, and then perform a model (in)validation step, as proposed in Steele and Vinnicombe (2001) to estimate the gap between the learned model and the true plant. This approach has the additional advantage that it can handle unstable plants. While learning coprime factors directly from data is an open problem, the results below suggest that, at least in the noiseless case, this can be accomplished by solving two convex Nevanlinna Pick interpolation problems.

Proposition 4. *Given input/output pairs* $\{r(z_i), y(z_i)\}_{i=1}^n$ *there exist stable transfer functions* $N(z), M(z)$ *such that* $y(z_i) = \frac{N(z_i)r(z_i)}{M(z_i)}$ *if and only if there exist* $u(z)$ *such that following conditions hold:*

$$P_N = \left[\frac{r(z_i)r^*(z_j) - u(z_i)u^*(z_j)}{1 - (z_i z_j^*)^{-1}} \right]_{i,j} \succeq 0$$
$$P_M = \left[\frac{y(z_i)y^*(z_j) - u(z_i)u^*(z_j)}{1 - (z_i z_j^*)^{-1}} \right]_{i,j} \succeq 0$$

Using Schur complements, these conditions can be transformed into convex LMIs in u. Once the Pick matrices P_N and P_M have been found, state space realizations for N and M can be obtained using for instance the formulas in Parrilo et al. (1998).

1.6 Conclusions

This extended abstract illustrates the challenges entailed in using ML to control dynamical systems. As shown here, learning stabilizing controllers places additional constraints on the architectures and on the algorithms used to train them. Thus, we argue that control-agnostic ML is unlikely to succeed in controlling challenging systems. Rather, the choice of representations and training has to be guided by systems theory.

REFERENCES

Cybenko, G. (1989). Approximation by superposition of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 303–314.

Doyle, J.C., Francis, B.A., and Tannenbaum, A. (1992). *Feedback Control Theory*. Macmillan, Toronto.

Khaneja, N. and Brockett, R. (1980). Dynamic feedback stabilization of nonholonomic systems. In *Proc. IEEE CDC, Phoenix, Dec.1999*, 1640–1645.

Parrilo, P.A., Sznaier, M., Sanchez-Pena, R.S., and Inanc, T. (1998). Mixed time/frequency-domain based robust identification. *Automatica*, 34(11), 1375–1389.

Savkin, A. and Petersen, I. (1997). A method for simultaneous strong stabilization of linear time-varying systems. *IFAC Proceedings Volumes*, 30(16), 165–169.

Sontag, E. (1992). Feedback stabilization using two-hidden-layer nets. *IEEE Trans. Neural Networks*, 3, 981–990.

Sontag, E. and Sussmann, H. (1980). Remarks on continuous feedback. In *Proc. IEEE CDC, Albuquerque, Dec.1980*, 916–921.

Steele, J. and Vinnicombe, G. (2001). Closed-loop time-domain model validation in the nu-gap metric. In *Proceedings of the 40th IEEE CDC*, volume 5, 4332–4337.

Zhou, K. and Doyle, J.C. (1998). *Essentials of Robust Control*. Prentice Hall.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On internally k-positive linear time-invariant system operators ⋆

Christian Grussler * **Thiago B. Burghi** ** **Somayeh Sojoudi** ***

* *Faculty of Mechanical Engineering, Technion – Israel Institute of Technology, Haifa, Israel (e-mail: cgrussler@technion.ac.il)*
** *Department of Engineering, University of Cambridge, Cambridge, United Kingdom (e-mail: tbb29@cam.ac.uk)*
*** *Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA, USA (e-mail: sojoudi@berkeley.edu)*

Abstract: Variation diminishment – the reduction in the number of sign changes and local extrema in a signal – is an intrinsic system property that lies at the heart of positive systems theory and over- and undershooting analysis in controlled systems. While, for general system operators, this property is difficult to verify, we show that it can be readily verified for the controllability and observability operators of finite-dimensional linear time-invariant systems under an internal k-positivity assumption. This complements earlier results on verifying this property for Hankel and Toeplitz operators, and establishes a bridge to internally positive systems theory. Our results provide a new framework for upper bounding the number of over- and undershoots in step responses, as well as a new realization theory of externally positive systems.

Keywords: positive systems, total positivity, k-positivity, variation diminishing, step response analysis

1. INTRODUCTION

Linear time-invariant (LTI) systems

$$\begin{aligned} x(t+1) &= Ax(t) + bu(t) \\ y(t) &= cx(t), \end{aligned} \tag{1}$$

that map nonnegative inputs u to nonnegative outputs y are characterized by a *nonnegative impulse* response $g(t) := cA^{t-1}b \geq 0$, $t \geq 1$ and are referred to as *externally positive* (Farina and Rinaldi, 2000). A particular property of such systems is the monotonicity of their step response, which motivated several studies on the avoidance of over- and undershooting in closed-loop design (Grussler and Rantzer, 2021; Darbha, 2003; Phillips and Seborg, 1988). Indeed, the number of sign changes of g, denoted by $S(g)$ and also known as the *variation* of g, equals the number of local extrema in the step response (see Section 2.1 for precise definitions).

The first problem addressed in this work is the extension of the positivity framework towards establishing upper bounds on the number of over- and undershoots in the step response of non-externally positive systems. We consider single-input-single-output (SISO) discrete-time systems of the form (1). There exist many lower bounds for this problem (Damm and Muhirwa, 2014; Swaroop and Niemann, 1996; El-Khoury et al., 1993), but only few upper bounds (El-Khoury et al., 1993). In our proposed solution, we express the impulse response of (1) as $g(t) = (\mathcal{O}(A,c)b)(t)$, with the *observability operator* given by

$$(\mathcal{O}(A,c)x_0)(t) := cA^t x_0,\ x_0 \in \mathbb{R}^n,\ t \geq 0 \tag{2}$$

The main idea is to bound $\mathrm{S}(g)$ by deriving a computationally tractable certificate for the largest integer k such that $\mathrm{S}(g) = \mathrm{S}(\mathcal{O}(A,c)b) \leq \mathrm{S}(b)$ for all $\{b : \mathrm{S}(b) \leq k\}$. An (observability) operator verifying this property is said to be *k-variation diminishing*. Although there exists a rich literature on variation diminishing transformations (see, e.g, the monographs by Karlin (1968); Pinkus (2009); Fallat and Johnson (2011)), including many recent contributions to system and control theory (see, e.g., (Grussler and Sepulchre, 2022; Grussler et al., 2021; Margaliot and Sontag, 2019; Wu and Margaliot, 2021)), it is unknown how to efficiently verify this property for $\mathcal{O}(A,c)$. This is not surprising, as the verification of external positivity (i.e., 0-variation diminishment) is computationally challenging in general (Blondel and Portier, 2002).

We therefore build our framework around the celebrated subclass of *internally positive systems*, i.e., systems characterized by nonnegative system matrices A, b and c (Luenberger, 1979), for which external positivity is an immediate consequence. These systems have received considerable interest due to their simplifying, scalable analysis properties (Rantzer and Valcher, 2018; Tanaka and Langbort, 2011; Farina and Rinaldi, 2000), and much effort has been put into the computation of internally positive realizations (Ohta et al., 1984; Farina, 1996; van den Hof, 1997; Farina and Rinaldi, 2000; Benvenuti and Farina, 2004). In fact, if we can find a realization with nonnegative A and c, it is also trivial that $\mathcal{O}(A,c)$ is 0-variation diminishing. Here, similar conditions and generalized realizability results are derived for $k > 0$. Our approach recovers the results of

⋆ This work received support by grants from ONR and NSF as well as under the Advanced ERC Grant Agreement Switchlet n.670645 and by DGAPA-UNAM under the grant PAPIIT RA105518.

El-Khoury et al. (1993) in the special case that k can be chosen arbitrarily large, which is a consequence of El-Khoury et al. (1993) using early results on variation diminishing mappings.

The second problem addressed in this work is the establishment of generalized internally positive realizations. Those are externally positive state-space systems that allow negative elements in b, at the cost of more restrictive assumptions on (A, c). It is envisioned that this will lead to generalizations of recent scalable positive systems analysis methods.

Finally, this work *bridges* recent external (input-output) variation diminishing system theory (Grussler and Sepulchre, 2022; Grussler et al., 2021) with internal (unforced, state-space) developments (see, e.g., Margaliot and Sontag (2019); Wu and Margaliot (2021)). In particular, we arrive at a positive realization theory of systems with variation diminishing Hankel operators, where $k = 0$ corresponds to classical external vs. internal positivity. The former has only recently been characterized in (Grussler and Sepulchre, 2022) in terms of so-called k compound systems, where our results are equivalent to simultaneous internally positive realizability of these k systems.

2. PRELIMINARIES

In this section, we briefly introduce some concepts that are essential for our results.

2.1 Variation diminishing maps

The *variation* of a sequence or vector u is defined as the number of sign-changes in u, i.e.,

$$\mathrm{S}(u) := \sum_{i\geq 1} \mathbb{1}_{\mathbb{R}_{<0}}(\tilde{u}_i\tilde{u}_{i+1}), \quad \mathrm{S}(0) := 0$$

where $\tilde{u}$ is the vector resulting from deleting all zeros in u.

Definition 1. A linear map $u \mapsto Xu$ is said to be *order-preserving k-variation diminishing* (OVD_k), $k \in \mathbb{Z}_{\geq 0}$, if for all u with $\mathrm{S}(u) \leq k$ it holds that

i. $\mathrm{S}(Xu) \leq \mathrm{S}(u)$.
ii. The sign of the first non-zero elements in u and Xu coincide whenever $\mathrm{S}(u) = \mathrm{S}(Xu)$.

If the second item is dropped, then $u \mapsto Xu$ is called *k-variation diminishing* (VD_k). For brevity, we simply say X is (O)VD_k.

2.2 Matrix k-positivity and compound matrices

For generic k, *total positivity theory* (Karlin, 1968) provides algebraic conditions for the OVD_k property by means of compound matrices. To define these, let the i-th elements of the r-tuples in

$$\mathcal{I}_{n,r} := \{v = \{v_1, \dots, v_r\} \subset \mathbb{N} : 1 \leq v_1 < \cdots < v_r \leq n\}$$

be defined by *lexicographic ordering*. Then, the (i, j)-th entry of the *r-th multiplicative compound matrix* $X_{[r]} \in \mathbb{R}^{\binom{n}{r}\times\binom{m}{r}}$ of $X \in \mathbb{R}^{n\times m}$ is defined by $\det(X_{(I,J)})$, where I is the i-th and J is the j-th element in $\mathcal{I}_{n,r}$ and $\mathcal{I}_{m,r}$, respectively. For example, if $X \in \mathbb{R}^{3\times 3}$, then $X_{[2]}$ reads

$$\begin{pmatrix} \det(X_{\{1,2\},\{1,2\}}) & \det(X_{\{1,2\},\{1,3\}}) & \det(X_{\{1,2\},\{2,3\}}) \\ \det(X_{\{1,3\},\{1,2\}}) & \det(X_{\{1,3\},\{1,3\}}) & \det(X_{\{1,3\},\{2,3\}}) \\ \det(X_{\{2,3\},\{1,2\}}) & \det(X_{\{2,3\},\{1,3\}}) & \det(X_{\{2,3\},\{2,3\}}) \end{pmatrix}.$$

Notice a nonnegative matrix verifies $X_{[1]} = X \geq 0$, which is equivalent to X being OVD_0. This can be generalized through the compound matrix as follows (Grussler and Sepulchre, 2022, Proposition 4).

Definition 2. Let $X \in \mathbb{R}^{n\times m}$ and $k \leq \min\{m, n\}$. X is called *k-positive* if $X_{[j]} \geq 0$ for $1 \leq j \leq k$, and *strictly k-positive* if $X_{[j]} > 0$ for $1 \leq j \leq k$. In case $k = \min\{m, n\}$, X is called *(strictly) totally positive*.

Proposition 3. Let $X \in \mathbb{R}^{n\times m}$ with $n \geq m$. Then, X is k-positive with $1 \leq k \leq m$ if and only if X is OVD_{k-1}.

The following is an important property of compound matrices (Horn and Johnson, 2012, Subsection 0.8.1).

Lemma 4. Let $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{p\times m}$. Then

$$(XY)_{[r]} = X_{[r]}Y_{[r]}$$

2.3 Hankel k-positivity and compound systems

The OVD_k property of LTI systems (1) has been studied by Grussler and Sepulchre (2022), where a distinction is made between LTI systems with OVD_k Toeplitz and Hankel operators. The latter are particularly relevant to this work. For $t \geq 0$, the *Hankel operator*

$$(\mathcal{H}_g u)(t) := \sum_{\tau=-\infty}^{-1} g(t-\tau)u(\tau) = \sum_{\tau=1}^{\infty} g(t+\tau)u(-\tau) \quad (3)$$

describes the evolution of the output y of an LTI system subjected to an input $u(t) = u(t)(1 - s(t))$, where $s(t)$ denotes the Heaviside function. The Hankel operator obeys the factorization

$$\mathcal{H}_g u = \mathcal{O}(A, c)(\mathcal{C}(A, b)u) \quad (4)$$

where

$$x(0) = \mathcal{C}(A, b)u := \sum_{\tau=-\infty}^{-1} A^{-\tau-1}bu(\tau),\ u \in \ell_\infty \quad (5)$$

denotes the *controllability operator*. Throughout this work, A is asymptotically stable, which implies that our operators are well-defined. For $t, j \in \mathbb{Z}_{>0}$, we define the *Hankel matrix*

$$H_g(t, j) := \begin{pmatrix} g(t) & g(t+1) & \dots & g(t+j-1) \\ g(t+1) & g(t+2) & \dots & g(t+j) \\ \vdots & \vdots & \ddots & \vdots \\ g(t+j-1) & g(t+j) & \dots & g(t+2(j-1)) \end{pmatrix} \quad (6a)$$

$$= \mathcal{O}^j(A, c)A^{t-1}\mathcal{C}^j(A, b) \quad (6b)$$

where

$$\mathcal{C}^j(A, b) := \begin{pmatrix} b & Ab & \dots & A^{j-1}b \end{pmatrix} \quad (6c)$$

$$\mathcal{O}^j(A, c) := \mathcal{C}^j(A^\mathsf{T}, c^\mathsf{T})^\mathsf{T}. \quad (6d)$$

Hankel k-positivity is defined as follows.

Definition 5. A system $G(z)$ is called *Hankel k-positive* if $\mathcal{H}_g$ is OVD_{k-1} ($k \geq 1$). If $k = \infty$, $G(z)$ is said to be *Hankel totally positive*.

Notice that Hankel 1-positivity coincides with the familiar property of external positivity (Farina and Rinaldi,

2000). A characterization of Hankel k-positivity is given by (Grussler and Sepulchre, 2022, Lemma 2):

Lemma 6. A system $G(z)$ is Hankel k-positive if and only if for all $j \in \mathbb{Z}_{\geq k}$, $H_g(1,j)$ is k-positive.

Using basic results from total positivity theory (Fallat et al., 2017), it is easy to show that k-positivity of Hankel matrices only requires checking the nonnegativity of consecutive minors. From (6b), each of these consecutive minors is given by

$$g_{[j]}(t) := \det(H_g(t,j)),$$

which is interpreted as the impulse response of an LTI system $G_{[j]}(z)$, called the *j-th compound system*. If (A,b,c) is a realization of $G(z)$, then $G_{[j]}(z)$ can be realized as

$$(A_{[j]}, \mathcal{C}^j(A,b)_{[j]}, \mathcal{O}^j(A,c)_{[j]}). \tag{7}$$

since

$$\det(H_g(t,j)) = H_g(t,j)_{[j]} = \mathcal{O}^j(A,c)_{[j]}(A_{[j]})^{t-1}\mathcal{C}^j(A,b)_{[j]}$$

by (6b) and Lemma 4.

3. INTERNALLY K-POSITIVE SYSTEM OPERATORS

In this section, we present our main results on the characterization of OVD_k observability and controllability operators to (1). The proofs of these results can be found in our full paper (Grussler et al., 2022), where an extension of the results to continuous-time is also studied.

We start with following simple lemma.

Lemma 7. For (A,b,c), the following are equivalent:

i. $\mathcal{C}(A,b)$ and $\mathcal{O}(A,c)$ are OVD_{k-1}, respectively.
ii. For all $t \geq k$, $\mathcal{C}^t(A,b)$ and $\mathcal{O}^t(A,c)$ are k-positive, respectively.

Notice that for $k = 1$ Lemma 7 recovers the well-known characterization of internal positivity in terms of the nonnegativity of (A,b), and (A,c), respectively. Next, we seek a finite-dimensional and computationally tractable characterization as our first main result.

Theorem 8. Let (A,b,c) be a realization of $G(z)$ such that A is k-positive. The following hold:

i. If $\mathcal{C}^j(A,b)_{[j]} \geq 0$ for $1 \leq j \leq k$, then $\mathcal{C}^t(A,b)$ is k-positive for all $t \geq k$.
ii. If $\mathcal{O}^j(A,c)_{[j]} \geq 0$ for $1 \leq j \leq k$, then $\mathcal{O}^t(A,c)$ is k-positive for all $t \geq k$.

The proof of this result involves extending a test for matrix k-positivity based on verifying the positivity of consecutive minors, see (Grussler et al., 2022, Theorem 3.6).

3.1 Impulse response analysis

For LTI systems, the total number of over- and undershoots (in the step response) equals the number of sign changes in the impulse response. Since $g(t) = (\mathcal{O}(A,c)b)(t)$, $\mathcal{O}(A,c)$ being OVD_{k-1} implies that response of (A,b,c) changes its sign at most $\mathrm{S}(b)$ times for all $\mathrm{S}(b) \leq k-1$, and has the same sign-changing order as b in case of an equal number of sign-changes. Since our framework is realization-dependent, we derive

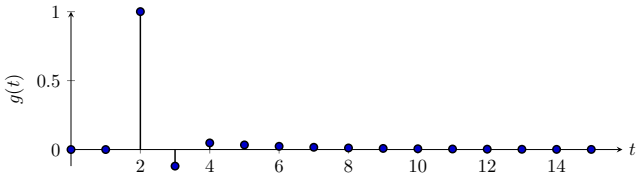


Fig. 1. The impulse response of (8) has two zero crossings, which coincides with our derived upper bound.

the following result, which is similar in spirit to classical positive realization theory (Ohta et al., 1984).

Theorem 9. Let (A,b,c) be a minimal realization of $G(z)$. Then, $G(z)$ admits a realization (A_+,b_+,c_+) such that A_+ and $\mathcal{O}(A_+,c_+)$ are OVD_{k-1} if and only if there exist a k-positive $N \in \mathbb{R}^{K\times K}$ and a $P \in \mathbb{R}^{n\times K}$ with $K \geq n$ such that $AP = PN$ and $\mathcal{O}^j(A,c)_{[j]}^{\mathsf{T}} \in \mathrm{cone}(P_{[j]})^*$ for $1 \leq j \leq k$, where $k \leq n$.

As a consequence of the order preservation in the OVD_k definition, this also provides a first simple result on generalized internally positive realizations.

Theorem 10. Let (A,b,c) be such that $\mathcal{O}(A,c)$ is OVD_1, $\mathrm{S}(b) \leq 1$. Further, assume that the first non-zero element in b is negative and the first non-zero element of the impulse response of (A,b,c) is positive. Then, (A,b,c) is externally positive.

Theorem 10 is a new contribution that has not been published in (Grussler et al., 2022).

Example Consider the following system, previously shown as an example in (El-Khoury et al., 1993):

$$G(z) = \frac{(z-0.22)(z-0.6)}{z^3(z-0.7)}$$

The transfer function $G(z)$ has a realization given by

$$A = \begin{pmatrix} 0.7 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad b = \begin{pmatrix} 0 \\ 1 \\ -0.82 \\ 0.132 \end{pmatrix} \quad c = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}^{\mathsf{T}} \tag{8}$$

It can be verified that this realization has totally positive A and $\mathcal{O}^4(A,c)$. By Lemma 7 and Theorem 8, $\mathcal{O}(A,c)$ is totally positive and the number of sign changes in the impulse response of $G(z)$ (and, hence, the number of extrema in its step response) is upper bounded by $\mathrm{S}(b) = 2$; the same upper bound was previously obtained by (El-Khoury et al., 1993). Figure 1 shows that this bound is tight. However, in contrast to (El-Khoury et al., 1993), our framework does not assume real poles or zeros. In particular, the modified transfer function

$$G_m(z) = \frac{(z-0.5+i)(z-0.5-i)}{z^3(z-0.7)},$$

can be realized with the same A and c as in (8) and with $b = (0\ 1\ -1\ 1.25)^{\mathsf{T}}$. This again provides a tight upper bound on the variation of the impulse response.

Finally, let A be as in (8), but $b = (-1\ 1\ 1\ 1)^{\mathsf{T}}$ and $c = (1\ 1\ 0\ 0)$. By Lemma 7 and Theorem 8, $\mathcal{O}(A,c)$ is OVD_1 (but not OVD_2) and $(g(1)\ g(2)) = (0\ 1.3)$. Applying Theorem 10 proves that (A,b,c) is externally positive without requiring a fully internally positive realization.

3.2 Internally Hankel k-positive systems

Next, we introduce the following subclass of Hankel k-positive systems:

Definition 11. (A, b, c) is called *internally Hankel k-positive* if A, $\mathcal{C}(A, b)$, and $\mathcal{O}(A, c)$ are OVD_{k-1} $(1 \leq k \leq n)$. If $k = n$, we say that (A, b, c) is *internally Hankel totally positive*.

Internally Hankel k-positive systems are thus OVD_{k-1} from past input u to $x(0)$, and from $x(0)$ to all future $x(t)$ and future output y. In particular, by (4), all internally Hankel k-positive systems are also (externally) Hankel k-positive, and setting $u \equiv 0$ recovers the k-positive property of unforced systems as partially studied in (Margaliot and Sontag, 2019; Wu and Margaliot, 2021). Thus, Definition 11 bridges the external and the unforced notions of variation diminishing LTI systems. A combination of Lemma 7 and Theorem 8 gives the following characterization of internal Hankel k-positivity.

Theorem 12. (A, b, c) is internally Hankel k-positive if and only if the realizations of the first k compound systems of (A, b, c) in (7) are (simultaneously) internally positive.

Similar to Theorem 9, we need to address the question of the existence of (minimal) Hankel k-positive realizations.

Theorem 13. A system $G(z)$ with order n and minimal realization (A, b, c) has a minimal internally Hankel k-positive realization, $k \leq n$, if and only if there exists a $P \in \mathbb{R}^{n \times n}$ with $\mathrm{rank}(P) = n$ such that for all $1 \leq j \leq k$

$$AP = PN \text{ for some } k\text{-positive } N \tag{9a}$$

$$\mathcal{C}^j(A, b)_{[j]} \in \mathrm{cone}(P_{[j]}), \tag{9b}$$

$$\mathcal{O}^j(A, c)_{[j]}^{\mathsf{T}} \in \mathrm{cone}(P_{[j]})^*. \tag{9c}$$

Finally, under an irreducibility condition, all unforced k-positive systems give rise to an internally Hankel k-positive system:

Proposition 14. Let $A \in \mathbb{R}^{n \times n}$ be k-positive with irreducible $A_{[j]}$, $1 \leq j \leq k$. Then there exists a $b \in \mathbb{R}^n$ such that $\mathcal{C}^j(A, b)_{[j]} > 0$ for all $1 \leq j \leq k$ and (A, b) is controllable.

REFERENCES

Benvenuti, L. and Farina, L. (2004). A tutorial on the positive realization problem. *IEEE Transactions on Automatic Control*, 49(5), 651–664.

Blondel, V.D. and Portier, N. (2002). The presence of a zero in an integer linear recurrent sequence is NP-hard to decide. *Linear Algebra and its Applications*, 351, 91 – 98.

Damm, T. and Muhirwa, L.N. (2014). Zero crossings, overshoot and initial undershoot in the step and impulse responses of linear systems. *IEEE Transactions on Automatic Control*, 59(7), 1925–1929.

Darbha, S. (2003). On the synthesis of controllers for continuous time lti systems that achieve a non-negative impulse response. *Automatica*, 39(1), 159 – 165.

El-Khoury, M., Crisalle, O.D., and Longchamp, R. (1993). Discrete transfer-function zeros and step-response extrema. *IFAC Proceedings Volumes*, 26(2, Part 2), 537 – 542.

Fallat, S., Johnson, C.R., and Sokal, A.D. (2017). Total positivity of sums, hadamard products and hadamard powers: Results and counterexamples. *Linear Algebra and its Applications*, 520, 242 – 259.

Fallat, S. and Johnson, C. (2011). *Totally Nonnegative Matrices*. Princeton University Press.

Farina, L. and Rinaldi, S. (2000). *Positive linear systems: theory and applications*. Pure and applied mathematics (John Wiley & Sons). Wiley.

Farina, L. (1996). On the existence of a positive realization. *Systems & Control Letters*, 28(4), 219 – 226.

Grussler, C., Damm, T., and Sepulchre, R. (2021). Balanced truncation of k-positive systems. *IEEE Transactions on Automatic Control*, 67, 526–531.

Grussler, C., Burghi, T.B., and Sojoudi, S. (2022). Internally Hankel k-positive systems. *SIAM Journal on Control and Optimization*. In print. Preprint at `arxiv.org/abs/2103.06962`.

Grussler, C. and Rantzer, A. (2021). On second-order cone positive systems. *SIAM Journal on Control and Optimization*, 59(4), 2717–2739.

Grussler, C. and Sepulchre, R. (2022). Variation diminishing linear time-invariant systems. *Automatica*, 136, 109985.

Horn, R.A. and Johnson, C.R. (2012). *Matrix Analysis*. Cambridge University Press, 2 edition.

Karlin, S. (1968). *Total positivity*, volume 1. Stanford University Press.

Luenberger, D. (1979). *Introduction to Dynamic Systems: Theory, Models & Applications*. John Wiley & Sons.

Margaliot, M. and Sontag, E.D. (2019). Revisiting totally positive differential systems: A tutorial and new results. *Automatica*, 101, 1–14.

Ohta, Y., Maeda, H., and Kodama, S. (1984). Reachability, observability, and realizability of continuous-time positive systems. *SIAM Journal on Control and Optimization*, 22(2), 171–180.

Phillips, S.F. and Seborg, D.E. (1988). Conditions that guarantee no overshoot for linear systems. *International Journal of Control*, 47(4), 1043–1059.

Pinkus, A. (2009). *Totally Positive Matrices*. Cambridge Tracts in Mathematics. Cambridge University Press.

Rantzer, A. and Valcher, M.E. (2018). A tutorial on positive systems and large scale control. In *2018 IEEE Conference on Decision and Control (CDC)*, 3686–3697.

Swaroop, D. and Niemann, D. (1996). Some new results on the oscillatory behavior of impulse and step responses for linear time invariant systems. In *Proceedings of 35th IEEE Conference on Decision and Control*, volume 3, 2511–2512 vol.3.

Tanaka, T. and Langbort, C. (2011). The bounded real lemma for internally positive systems and h-infinity structured static state feedback. *IEEE Transactions on Automatic Control*, 56(9), 2218–2223.

van den Hof, J. (1997). Realization of positive linear systems. *Linear Algebra and its Applications*, 256, 287–308.

Wu, C. and Margaliot, M. (2021). Diagonal stability of discrete-time k-positive linear systems with applications to nonlinear systems. *IEEE Transactions on Automatic Control*, 1–1.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Approximation of determistic mean field games with control-affine dynamics

Justina Gianatti * **Francisco J. Silva Álvarez** **

* *CIFASIS-CONICET-UNR, Bv. 27 de Febrero 210 bis, Rosario, S2000EZP Argentina (e-mail: gianatti@cifasis-conicet.gov.ar)*
** *Institut de recherche XLIM-DMI, UMR-CNRS 7252, Faculté des Sciences et Techniques, Université de Limoges, Limoges, CO 87060 France (e-mail: francisco.silva@unilim.fr)*

Keywords: Mean field games, Hamilton-Jacobi-Bellman equations, fictitious play method, convergent scheme, numerical examples.

AMS subject classifications: 35F35, 35Q91, 35A35.

1. EXTENDED ABSTRACT

Mean Field Games (MFGs) systems were introduced independently by Huang et al. (2006) and Lasry and Lions (2007) in order to model dynamic differential games with a large number of *indistinguishable small players*. In the model proposed in Lasry and Lions (2007), the asymptotic equilibrium is described by means of a system of two Partial Differential Equations (PDEs). The first equation, together with a final condition, is a Hamilton-Jacobi-Bellman (HJB) equation describing the value function of a *typical player* whose cost function depends on the distribution m of the entire population. The second equation is a Fokker-Planck (FP) equation which, together with an initial distribution m_0, describes the fact that m evolves following the optimal dynamics of the typical player.

In the case where both equations have a nondegenerate second order term, several numerical methods have been proposed to approximate solutions to the MFG system. In this context, convergent finite difference and Semi-Lagrangian (SL) schemes have been proposed in Achdou and Capuzzo-Dolcetta (2010); Achdou et al. (2013), and Carlini and Silva (2018) respectively.

In the case where both equations in the MFG system have no second order terms, we say that the MFG system is of first order. This system characterizes the limit behavior of equilibria of *deterministic* and *symmetric* differential games as the number of players tends to infinity. In this framework, a SL scheme has been proposed in Carlini and Silva (2014), where a convergence result is established when the state dimension is equal to one. We also refer the reader to Chowdhury et al. (2021) for a recent extension of this result to the case where non-local and fractional diffusion terms appear in both equations of the MFG system. To the best of our knowledge, the only convergence result in general state dimension of an approximation of the first order MFG system has been established in Hadikhanloo and Silva (2019). In these aforementioned works, it is supposed that a typical agent in the game controls directly its velocity and minimizes a cost functional whose dependence on the state variable is rather restrictive.

In this talk, we consider deterministic MFGs where the dynamics of a typical agent is non-linear with respect to the state variable and affine with respect to the control variable. Particular instances of this problem are MFGs where the typical agent controls its acceleration (see e.g. Achdou et al. (2020); Cannarsa and Mendico (2020)). We propose a fully discrete scheme that combines a semi-Lagrangian type discretization of the HJB equation, which takes advantage of the particular structure of the dynamics, and a discretization of the FP equation which has a probabilistic interpretation in terms of an underlying discrete time and finite state Markov chain X_n. The scheme that we propose takes then the form of a discrete time an finite state MFG (see Gomes et al. (2010)) which can be solved by the *fictitious play method* (see Hadikhanloo and Silva (2019)). Our main result is the convergence of solutions to this approximation towards MFG equilibria, which is numerically illustrated by several examples dealing with MFGs with control on the acceleration.

REFERENCES

Achdou, Y., Camilli, F., and Capuzzo-Dolcetta, I. (2013). Mean field games: convergence of a finite difference method. *SIAM J. Numer. Anal.*, 51(5), 2585–2612.

Achdou, Y. and Capuzzo-Dolcetta, I. (2010). Mean field games: numerical methods. *SIAM J. Numer. Anal.*, 48(3), 1136–1162.

Achdou, Y., Mannucci, P., Marchi, C., and Tchou, N. (2020). Deterministic mean field games with control on the acceleration. *NoDEA Nonlinear Differential Equations Appl.*, 27(3), Paper No. 33, 32.

Cannarsa, P. and Mendico, C. (2020). Mild and weak solutions of mean field game problems for linear control systems. *Minimax Theory Appl.*, 5(2), 221–250.

Carlini, E. and Silva, F.J. (2014). A fully discrete semi-Lagrangian scheme for a first order mean field game problem. *SIAM J. Numer. Anal.*, 52(1), 45–67.

Carlini, E. and Silva, F.J. (2018). On the discretization of some nonlinear Fokker-Planck-Kolmogorov equations and applications. *SIAM J. Numer. Anal.*, 56(4), 2148–2177.

Chowdhury, I., Ersland, O., and Jakobsen, E.R. (2021). On numerical approximations of fractional and nonlocal mean field games. Preprint.

Gomes, D.A., Mohr, J., and Souza, R. (2010). Discrete time, finite state space mean field games,. *Journal de Mathématiques Pures et Appliquées*, 93, 308–328.

Hadikhanloo, S. and Silva, F.J. (2019). Finite mean field games: fictitious play and convergence to a first order continuous mean field game. *J. Math. Pures Appl. (9)*, 132, 369–397.

Huang, M., Malhamé, R.P., and Caines, P.E. (2006). Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.*, 6(3), 221–251.

Lasry, J.M. and Lions, P.L. (2007). Mean field games. *Jpn. J. Math.*, 2, 229–260.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Lossy Schrödinger Bridges: The most likely transport between unbalanced marginals

Yongxin Chen * **Tryphon T. Georgiou** ** **Michele Pavon** ***

* *School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA (e-mail: yongchen@gatech.edu).*
** *Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA 92697, USA (e-mail: tryphon@uci.edu).*
*** *Department of Mathematics "Tullio Levi-Civita", Università di Padova, 35121 Padova, Italy (e-mail: pavon@math.unipd.it).*

Abstract: The problem to reconcile observed marginal distributions with a given prior was posed by E. Schrödinger in 1932/32, and is now known as the Schroödinger Bridge Problem. It represents a stochastic counterpart of the Optimal Mass Transport (OMT). In either setting, the problem to interpolate between "unbalanced" marginals has been approached by introducing source/sink terms into the transport equations, in an adhoc manner, chiefly driven by applications in image registration. In the present work we developed a formalism to interpolate between "unbalanced" marginals in the original spirit of E. Schrödinger, seeking the most likely transport of particles that may vanish along their path between given end points in time. In this, we develop a Schrödinger system of equations that accounts for losses, by allowing particles to "jump" into a coffin state according to a suitable probabilistic law. The solution of the Schrödinger system allows constructing a stochastic evolution that reconciles the given unbalanced marginals.

Keywords: Estimation, Stochastic control, Large deviation theory.

1. EXTENDED SUMMARY

The present extended abstract is based on Chen et al. (2021). The main contribution is an extension of the paradigm of the Schroödinger Bridge Problem that seeks to reconcile marginal distributions with a prior probability law, to the case where the marginal distributions are unbalanced, corresponding to differing total mass that needs to be accounted for via losses. A physical instance that provides motivation for the scenario that we envision can be described as follows.

Consider the problem of estimating the velocity field of ocean currents by releasing into the water a cloud of tracer particles and by sampling their distribution at a later time. The diffusion coefficient is assumed known and the original cloud that is released at time $t = 0$ consists of N particles. These are expected to remain in suspension for a duration of time while they diffuse and drift with the current. At time $t = 1$, their distribution is sampled again. Some of the particles in the meantime have sunk, so that the number of found particles is less than N. Suppose this experiment is performed several times, treating the model originating from previous experiments as a "prior". Is it conceivable to "improve" a prior model in a rational way? More explicitly, by relying on a prior model and the new sampling result, is it possible to determine an updated model that represents the most probable way that the tracer cloud may have been transported?

Further motivation is provided by natural processes that involve micro-organisms that may reproduce or die out along the way between measurements, sediment or pollution transport processes that dissipate or accentuate between observations, or virtual processes where fusion of data or morphing of images call for interpolation between distributions of varying total mass. However, our basic framework appeals to a setting where the mechanism generating losses has a physical origin as highlighted in the example with tracer particles.

At first sight, this problem appears to be of a different nature than those treated in the theory of Large Deviations Varadhan (1966, 1984); Dembo and Zeitouni (2009), in that the sought path-space measure is not a probability measure per se. Nevertheless, in spite of the paucity of the available data, it is possible to solve this inverse problem by a natural embedding technique. A byproduct is a physically motivated framework to interpolate distributions of unequal mass (integrals). The blueprint for the rationale in our work is the celebrated duo of papers by E. Schrödinger in 1931/32 Schrödinger (1931, 1932) where he considered the problem of reconciling (equal-mass) marginal distributions with a prior stochastic evolution.

In our formulation of the unbalanced Schrödinger Bridge Problem (uSBP), the marginals cannot be assumed to be probability distributions as their integrals differ due to losses. To this end, we embed the distributions into a frame that includes a coffin/extinction state, leading to a probability law on a continuum together with a discrete state. Thereupon, we find the updated law and killing rate that minimize the relative entropy to the prior with losses, and are consistent with the two marginals. In the special

case when the marginals are already consistent with the prior, naturally, the solution coincides with the lossy prior, differently from what happens in other formulations of SBP with killing which are based on Feyman-Kac functionals Nagasawa (1990); Wakolbinger (1989); Blaquière (1992); Dawson et al. (1990); Aebi and Nagasawa (1992); Léonard (2011); Chen et al. (2015, 2017), and unbalanced transport Chizat et al. (2018a,b); Chen et al. (2019); Koehl et al. (2021).

We would like to stress the fact that earlier proposals on how to address unbalanced transport Chen et al. (2015); Georgiou et al. (2008); Jiang et al. (2011); Chizat et al. (2018a,b); Chen et al. (2019); Koehl et al. (2021) resorted to ad-hoc regularization to cope with the unequal marginals, where either adjustment at the two ends or the presence of a source/sink contribution dynamically adjusts the differing end-point masses. In contrast, our formulation is cast in terms of prior knowledge and large-deviation theory, advancing a viewpoint that is close in spirit to the original rationale of E. Schrödinger. To this end, we consider below a diffusion process with killing and seek the closest update of the corresponding law that is in agreement with the marginal data. That is, we seek the most likely diffusion process with a suitably updated killing that matches the two marginals of unequal mass. We next summarize the main technical statements of the theory that we will present.

Consider a diffusion process

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{1}$$

over the Euclidean space $\mathbb{R}^n$, with a nonnegative killing rate $V(t, x)$ (assume $V(\cdot, \cdot)$ is jointly continuous with respect to (t, x) and not constantly zero). A thought experiment similar to Schrödinger's, calls for a large number N of trajectories over a time interval $[0, 1]$, that are independently sampled from (1) with initial probability distribution ρ_0, and a recorded empirical distribution for the surviving particles at time $t = 1$ approximated by ρ_1, which is inconsistent with the prior law, that is,

$$\rho_1(\cdot) \neq \int_{\mathbb{R}^n} q(0, x, 1, \cdot)\rho_0(x)dx.$$

The kernel $q(0, x, 1, y)$ is no longer a probability kernel in that

$$\int_{\mathbb{R}^n} q(0, x, 1, y)dy \neq 1,$$

in general, and thus, neither $\int_{\mathbb{R}^n} q(0, x, 1, \cdot)\rho_0(x)dx$ nor ρ_1 are necessarily probability densities, due to killing. In particular, $\int \rho_1(x)dx = N_s/N \leq 1$ where N_s denotes the number of survival particles at time 1. Just as in the standard SBP, we consider continuous distributions, assuming that N is large, and seek to identify the most likely behavior of the particles. By behavior we mean the most likely evolution of the particles along with the most likely times that the particles may have gotten killed (or, absorbed by an underlying medium).

As in the standard Schrödinger bridge, the problem arising from the above thought experiment can be formally stated using the theory of *large deviations* Dembo and Zeitouni (2009). However, in this case, the space of trajectories needs to be modified to accommodate for possible killing of particles. To this end, we augment the state space of the diffusion $\mathbb{R}^n$ with a "coffin state" $\mathfrak{c} \notin \mathbb{R}^n$, see (Øksendal, 2000, Subsection 8.2), resulting in the state space

$$\mathcal{X} = \mathbb{R}^n \cup \{\mathfrak{c}\}.$$

Let $\Omega = D([0, 1], \mathcal{X})$ be the Skorokhod space over $\mathcal{X}$, that is, each element in Ω is a càdlàg over $\mathcal{X}$ (Billingsley, 1999, page 121). Denote by $\mathcal{P}(\Omega)$ and $\mathcal{P}(\mathcal{X})$ the spaces of probability distributions over Ω and $\mathcal{X}$, respectively. Each diffusion process X_t $(t \in [0, 1])$ on $\mathbb{R}^n$ with killing corresponds to a process $\mathbf{X}_t$ taking values in $\mathcal{X}$, and therby, to a law in $\mathcal{P}(\Omega)$.

In our unbalanced SBP setting, the set of probability laws over path space $\mathcal{P}(\Omega)$ that are in alignment with the observations is

$$\{\mathbf{P} \in \mathcal{P}(\Omega) \mid \mathbf{P}_0 = p_0,\ \mathbf{P}_1 = p_1\},$$

where p_0, p_1 are the natural augmentation of ρ_0, ρ_1 so that they belong in $\mathcal{P}(\mathcal{X})$, respectively. Specifically, assuming that $\int_{\mathbb{R}^n} \rho_1(x)dx \leq 1$, we set

$$p_0 = (\rho_0(\cdot), 0) \tag{2a}$$

and

$$p_1 = (\rho_1(\cdot), 1 - \int_{\mathbb{R}^n} \rho_1(x)dx), \tag{2b}$$

and arrive at the following.

Unbalanced Schrödinger Bridge Problem (uSBP): Determine

$$\mathbf{P}^\star := \arg \min_{\mathbf{P} \in \mathcal{P}(\Omega)} \{\mathbb{D}(\mathbf{P}\|\mathbf{R}) \ \mid \ \mathbf{P}_0 = p_0, \mathbf{P}_1 = p_1\}. \tag{3}$$

Here, with the notation $\mathbf{R}(\cdot) = \int_{\mathcal{X}^2} \mathbf{R}^{xy}(\cdot)\mathbf{R}_{01}(dxdy)$, and $\mathbf{P}(\cdot) = \int_{\mathcal{X}^2} \mathbf{P}^{xy}(\cdot)\mathbf{P}_{01}(dxdy)$ we denote path measures, whereas $\mathbf{R}_{01}$ $(\mathbf{P}_{01})$ denotes the joint marginal distribution of $\mathbf{R}$ $(\mathbf{P})$ over the marginal $\mathbf{X}_{0,1}$, and $\mathbf{R}^{xy}$ $(\mathbf{P}^{xy})$ denotes the law conditioned on $\mathbf{X}_0 = x \in \mathcal{X}$ and $\mathbf{X}_1 = y \in \mathcal{X}$. Then, also, $\mathbb{D}(\mathbf{P}\|\mathbf{R})$ denotes the Kullback-Leibler divergence. A relation with a static SBP emerges.

Static uSBP: Determine

$$\boldsymbol{\pi}^\star := \arg \min_{\boldsymbol{\pi} \in \mathcal{P}(\mathcal{X}^2)} \{\mathbb{D}(\boldsymbol{\pi}\|\mathbf{R}_{01}) \ \mid \ \boldsymbol{\pi}_0 = p_0, \boldsymbol{\pi}_1 = p_1\}. \tag{4}$$

The two formulations can be seen to be equivalent, as follows from the identity

$$\mathbb{D}(\mathbf{P}\|\mathbf{R}) = \mathbb{D}(\mathbf{P}_{01}\|\mathbf{R}_{01}) + \int_{\mathcal{X}^2} \mathbb{D}(\mathbf{P}^{xy}\|\mathbf{R}^{xy})\mathbf{P}_{01}(dxdy).$$

The equivalence is highlighted in the following theorem.

Theorem 1. (see Chen et al. (2021)). Suppose $\mathbf{P}^\star$ solves the dynamic uSBP (3), then $\mathbf{P}^\star_{01}$ also solves the static uSBP (4). On the other hand, if $\boldsymbol{\pi}^\star$ solves (4), then setting $\mathbf{P}^\star = \int_{\mathcal{X}^2} \mathbf{R}^{xy}(\cdot)\boldsymbol{\pi}^\star(dxdy)$ solves (3), while $\mathbf{P}^\star_{01} = \boldsymbol{\pi}^\star$.

The solutions to (3) and (4) are of the form

$$\mathbf{P}^\star = f(\mathbf{X}_0)g(\mathbf{X}_1)\mathbf{R}, \text{ and} \tag{5}$$

$$\boldsymbol{\pi}^\star = f(\mathbf{X}_0)g(\mathbf{X}_1)\mathbf{R}_{01}, \tag{6}$$

respectively. Explicit solution that includes an update on the prior killing rate as well as determining the change of measure effected by f, g, can be computed as follows.

The Fokker-Planck equation for a diffusion (1) with killing rate $V(t, x)$ is

$$\partial_t R_t + \nabla \cdot (bR_t) + VR_t = \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 (a_{ij} R_t)}{\partial x_i \partial x_j}. \tag{7}$$

Throughout we assume that $a(t,X)=\sigma(t,X)\sigma(t,X)'$ is a positive definite matrix. A corresponding Schrödinger system and its relation to the law of $\mathbf{P}^\star$ can be expressed after reparametrizing the pair (f,g) of functions on $\mathcal{X}$ as follows

$$f(x)\mathbf{R}_0(x)=\begin{cases}\hat\varphi(0,x) & \text{if } x\in\mathbb{R}^n\\ \hat\psi(0) & \text{if } x=\mathfrak{c},\end{cases} \tag{8a}$$

$$g(y)=\begin{cases}\varphi(1,y) & \text{if } y\in\mathbb{R}^n\\ \psi(1) & \text{if } y=\mathfrak{c}.\end{cases} \tag{8b}$$

A generalized Schrödinger system along with the nonlinear coupling constraints then takes the form

$$\partial_t\hat\varphi=-\nabla\cdot(b\hat\varphi)-V\hat\varphi+\frac{1}{2}\sum_{i,j=1}^n\frac{\partial^2(a_{ij}\hat\varphi)}{\partial x_i\partial x_j} \tag{9a}$$

$$\frac{d\hat\psi}{dt}=\int V\hat\varphi(t,x)dx \tag{9b}$$

$$\partial_t\varphi=-b\cdot\nabla\varphi+V\varphi-\frac{1}{2}\sum_{i,j=1}^n a_{ij}\frac{\partial^2\varphi}{\partial x_i\partial x_j}-V\varphi \tag{9c}$$

$$\frac{d\psi}{dt}=0 \tag{9d}$$

$$\rho_0=\varphi(0,\cdot)\hat\varphi(0,\cdot) \tag{9e}$$

$$\rho_1=\varphi(1,\cdot)\hat\varphi(1,\cdot) \tag{9f}$$

$$\hat\psi(0)=0 \tag{9g}$$

$$\psi(1)\hat\psi(1)=1-\int\rho_1. \tag{9h}$$

The solvability of this Schrödinger system is claimed in the next theorem.

Theorem 2. (see Chen et al. (2021)). Let R be the law of a diffusion (1) with nontrivial killing rate $V(t,x)$ and $a(t,x)=\sigma(t,x)\sigma(t,x)'$ being positive definite for all $(t,x)\in\mathbb{R}\times\mathbb{R}^n$, and assume that ρ_0,ρ_1 are absolutely continuous with respect to the Lebesgue measure. There exists a unique (up to a constant scaling) 4-tuple $(\hat\varphi(t,x),\hat\psi(t),\varphi(t,x),\psi(t))$ of non-negative functions that satisfies the Schrödinger system (9).

Discretization of the Schrödinger system (9) leads to an efficient algorithm to compute $(\hat\varphi(t,x),\hat\psi(t),\varphi(t,x),\psi(t))$, and thereby, $\mathbf{P}^\star$ as well as the corresponding Fokker-Planck equation for the corresponding marginals. This will be discussed in the talk and is detailed in Chen et al. (2021). Further, a dynamic formulation of the updated law is possible in the form of a corresponding diffusion process as we explain next.

We denote by P_t the marginal of $\mathbf{X}_t$ restricted to the first component in $\mathcal{X}$, and by q_t the probabilty of the coffin state. Thus, we use the vectorial notation

$$\mathbf{P}_t=:(P_t,q_t).$$

Accordingly, for the marginals $\mathbf{R}_t=(R_t,s_t)$ of the prior, R_t satisfies the Fokker-Planck equation (7) while $s_t=1-\int_{\mathbb{R}^n}R_t(x)dx$. The solution $\mathbf{P}^\star$ to (3) is then characterized by the following theorem.

Theorem 3. (see Chen et al. (2021)). The solution $\mathbf{P}^\star$ to (3) corresponds to a diffusion process

$$dX_t=(b(t,X_t)+a(t,X_t)\nabla\log\varphi(t,X_t))dt+\sigma(t,X_t)dW_t \tag{10}$$

with killing rate $\psi V/\varphi$, where φ is obtained from the solution of the generalized Schrödinger system (9). Accordingly,

$$\partial_tP_t+\nabla\cdot((b+a\nabla\log\varphi)P_t)=\frac{1}{2}\sum_{i,j=1}^n\frac{\partial^2(a_{ij}P_t)}{\partial x_i\partial x_j}-\frac{\psi}{\varphi}VP_t. \tag{11}$$

The original Schrödinger bridge problem, when there is no killing, is known to be equivalent to the stochastic control problem of minimizing control energy subject to the marginal two end-point constraints Chen et al. (2016), or equivalently, to a fluid dynamic formulation whereby the velocity field $u(t,\cdot)$ effecting the flow minimizes this action integral, namely,

$$\min_{P_t(\cdot),u(t,\cdot)}\int_0^1\int_{\mathbb{R}^n}\frac{1}{2}\|u(t,x)\|^2P_tdxdt \tag{12a}$$

$$\partial_tP_t+\nabla\cdot((b+\sigma u)P_t)-\frac{1}{2}\sum_{i,j=1}^n\frac{\partial^2(a_{ij}P_t)}{\partial x_i\partial x_j}=0 \tag{12b}$$

$$P_0=\rho_0,\quad P_1=\rho_1. \tag{12c}$$

The optimization takes place over the *feedback control policy-flow field* $u(t,x)$ together with the corresponding density flow $P_t(x)$. Below, we present an analogous formulation for the Schrödinger bridge problems with unbalanced marginals.

Along the flow, the killing rate may deviate from the prior V and is to be determined. To quantify the deviation of the posterior killing rate from the prior, we introduce an entropic cost inside the action integral, to penalize changes in the ratio $\alpha(t,x)$ between the posterior and the prior killing rate. That is, α is an added optimization variable which is $\alpha(t,x)\geq 0$, and with the posterior killing rate being αV. To penalize differences between the posterior and the prior killing rates we introduce the factor

$$\alpha\log\alpha-\alpha+1 \tag{13}$$

inside the action integral, which is convex and achieves the minimal value 0 at $\alpha=1$. This entropy cost has been used in Léonard (2014, 2016) to study Schrödinger bridge problem over graphs. It is associated with the large deviation principle for continuous-time Markov chain with discrete state. Combining this entropic cost term for the ratio of killing rates with (12) we arrive at

$$\min_{P,u,\alpha}\int_0^1\int_{\mathbb{R}^n}\left[\frac{1}{2}\|u(t,x)\|^2P_t+(\alpha\log\alpha-\alpha+1)VP_t\right]dxdt \tag{14a}$$

$$\partial_tP_t+\nabla\cdot((b+\sigma u)P_t)+\alpha VP_t-\frac{1}{2}\sum_{i,j=1}^n\frac{\partial^2(a_{ij}P_t)}{\partial x_i\partial x_j}=0 \tag{14b}$$

$$P_0=\rho_0,\quad P_1=\rho_1. \tag{14c}$$

Note that the control strategy has now two components, a drift term $u(t,x)$ and a correcting term $\alpha(t,x)$ for the killing rate. We conclude with the existence and form of solutions to (14).

Theorem 4. Let $(\hat{\varphi}(t,x), \hat{\psi}(t), \varphi(t,x), \psi(t))$ be the solution to the Schrödinger system (9), then the solution to (14) is given by the choice

$$u^\star(t,x) = \sigma(t,x)' \nabla \log \varphi(t,x) \tag{15a}$$

$$\alpha^\star(t,x) = \frac{\psi(t)}{\varphi(t,x)} \tag{15b}$$

$$P_t(x) = \varphi(t,x)\hat{\varphi}(t,x). \tag{15c}$$

In summary, the key contribution in this work is to address in a rigorous manner the needed update of prior evolution and killing mechanism, through large deviation theory, that extends Schrödinger's dictum of seeking the most likely law that reconciles given marginal distributions. The solution that emerges differs in an essential way from earlier approaches to do the same via Feynman-Kac multiplicative reweighing Nagasawa (1990); Wakolbinger (1989); Blaquière (1992); Dawson et al. (1990); Aebi and Nagasawa (1992); Léonard (2011); Chen et al. (2015, 2017). It is important to underscore that in these earlier approaches, when the prior is consistent with the given marginals, the solution fails to coincide with the prior as one would expect and want. Such a natural requirement for the solution is inherent in our large deviation formulation of the unbalanced Schrödinger Bridge Problem.

REFERENCES

Aebi, R. and Nagasawa, M. (1992). Large deviations and the propagation of chaos for Schrödinger processes. *Probability Theory and Related Fields*, 94(1), 53–68.

Billingsley, P. (1999). *Convergence of probability measures*. John Wiley & Sons.

Blaquière, A. (1992). Controllability of a Fokker-Planck equation, the Schrödinger system, and a related stochastic optimal control (revised version). *Dynamics and Control*, 2(3), 235–253.

Chen, Y., Georgiou, T., and Pavon, M. (2017). Optimal steering of a linear stochastic system to a final probability distribution, Part III. *arXiv:1608.03622, IEEE Trans. on Automatic Control, to appear*.

Chen, Y., Georgiou, T.T., and Pavon, M. (2015). Optimal steering of inertial particles diffusing anisotropically with losses. In *Proc. American Control Conf.*, 1252–1257.

Chen, Y., Georgiou, T.T., and Pavon, M. (2016). On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2), 671–691.

Chen, Y., Georgiou, T.T., and Pavon, M. (2021). The most likely evolution of diffusing and vanishing particles: Schrodinger bridges with unbalanced marginals. *arXiv preprint arXiv:2108.02879*.

Chen, Y., Georgiou, T.T., and Tannenbaum, A. (2019). Interpolation of matrices and matrix-valued densities: The unbalanced case. *European Journal of Applied Mathematics*, 30(3), 458–480.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.X. (2018a). Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314), 2563–2609.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.X. (2018b). Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11), 3090–3123.

Dawson, D., Gorostiza, L., and Wakolbinger, A. (1990). Schrödinger processes and large deviations. *Journal of mathematical physics*, 31(10), 2385–2388.

Dembo, A. and Zeitouni, O. (2009). *Large deviations techniques and applications*, volume 38. Springer Science & Business Media.

Georgiou, T.T., Karlsson, J., and Takyar, M.S. (2008). Metrics for power spectra: an axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3), 859–867.

Jiang, X., Luo, Z.Q., and Georgiou, T.T. (2011). Geometric methods for spectral analysis. *IEEE Transactions on Signal Processing*, 60(3), 1064–1074.

Koehl, P., Delarue, M., and Orland, H. (2021). Physics approach to the variable-mass optimal-transport problem. *Physical Review E*, 103(1), 012113.

Léonard, C. (2011). Stochastic derivatives and generalized h-transforms of Markov processes. *arXiv preprint arXiv:1102.3172*.

Léonard, C. (2014). A survey of the Schrödinger problem and some of its connections with optimal transport. *Dicrete Contin. Dyn. Syst. A*, 34(4), 1533–1574.

Léonard, C. (2016). Lazy random walks and optimal transport on graphs. *The annals of Probability*, 44(3), 1864–1915.

Nagasawa, M. (1990). Stochastic variational principle of Schrödinger processes. In *Seminar on Stochastic Processes, 1989*, 165–175. Springer.

Øksendal, B. (2000). *Stochastic differential equations: an introduction with applications, Fifth Edition*. Springer Science & Business Media.

Schrödinger, E. (1931). Über die Umkehrung der Naturgesetze. *Sitzungsberichte der Preuss Akad. Wissen. Phys. Math. Klasse, Sonderausgabe*, IX, 144–153.

Schrödinger, E. (1932). Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2(4), 269–310. Presses universitaires de France.

Varadhan, S.S. (1966). Asymptotic probabilities and differential equations. *Communications on Pure and Applied Mathematics*, 19(3), 261–286.

Varadhan, S.S. (1984). *Large deviations and applications*. SIAM.

Wakolbinger, A. (1989). A simplified variational characterization of Schrödinger processes. *Journal of mathematical physics*, 30(12), 2943–2946.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Designing and Combining Designs ⋆

Mario Osvin Pavčević * **Kristijan Tabak** **

* *University of Zagreb, Faculty of electrical engineering and computing, Department of applied mathematics, Croatia (e-mail: mario.pavcevic@fer.hr).*
** *Rochester Institute of Technology, Zagreb Campus, Croatia (e-mail: kristijan.tabak@croatia.rit.edu)*

Abstract: The fact that further constructions often don't bring breakthrough results motivates us to combine designs to be particles of other combinatorial structures. One of them are mosaics of designs, where instead of having a matrix presenting incidences of a design, one might fill the matrix with incidences of more than one design. Another way of combining designs are design cubes. They can be thought of as 3-dimensional incidence 0-1 matrices, such that each 2-dimensional incidence submatrix satisfies the properties of a design. In this paper we shall be concentrated on designing and combining t-designs although we are aware of the fact that the ideas presented here might work and be interesting for other sorts of combinatorial designs.

Keywords: t-design, mosaic of designs, design cube, difference set
AMS subject classification: 05B05, 05B30

1. INTRODUCTION AND PRELIMINARIES

Combinatorial designs are *set systems* (a collection of subsets of a given set) and consist of two types of objects, *points* (elements of the given set) and *blocks* (subsets of the given set, e.g. members of the given collection), satisfying some additional properties. There are different kinds of combinatorial designs; a good overview of most of them can be found in Colbourn et al. (2007). They are often represented with an *incidence matrix*, which is a 0-1 matrix with rows labelled by points and columns labelled by blocks, where its (i, j)-th entry indicates whether a point lies in a block or not.

In this paper we shall construct and combine t-designs, but it is worth mentioning that the main ideas of combining designs presented here would work for some other types of designs as well. A t-design with parameters (v, k, λ) is a collection $\mathcal{B}$ of k-element subsets (blocks) of a v-element set X (of points), such that every t-element subset of X is contained in exactly λ blocks. In such a case we speak sometimes of a t-(v, k, λ) design. It is known that t, v, k and λ must satisfy a number of more or less complicated (necessary) divisibility conditions, whereas for $t \geq 2$ the existence of a t-design with parameters (v, k, λ) is known only for particular cases which are not described by any general set of sufficient conditions. The number of blocks b can be calculated as $b = \lambda \cdot \binom{v}{t} / \binom{k}{t}$. A design is called *symmetric* if $b = v$. Each point of a t-design is incident with the same number of blocks, usually denoted by $r = \lambda \cdot \binom{v-1}{t-1} / \binom{k-1}{t-1}$.

Example 1. The most prominent and cited example in combinatorial design theory is the Fano plane, having parameters 2-$(7, 3, 1)$. Its incidence matrix can be chosen to be cyclic and given as

$$M = \begin{bmatrix} 0&1&1&0&1&0&0 \\ 0&0&1&1&0&1&0 \\ 0&0&0&1&1&0&1 \\ 1&0&0&0&1&1&0 \\ 0&1&0&0&0&1&1 \\ 1&0&1&0&0&0&1 \\ 1&1&0&1&0&0&0 \end{bmatrix}$$

Where are the borders of knowledge considering the existence of t-designs? For $t = 2$, it is not known whether 2-$(51, 6, 1)$, 2-$(40, 10, 3)$ or 2-$(39, 13, 6)$ exist. If you add the condition for the design to be symmetric, then the smallest examples with non-determined existence are 2-$(157, 13, 1)$ (the projective plane of order 12) or 2-$(81, 16, 3)$. The smallest 3-design with existence in question is 3-$(16, 7, 5)$ and the smallest 4-design 4-$(13, 6, 6)$. Let us point out that at the first glance these parameter sets look rather small.

2. CONSTRUCTION METHODS

A design is given when the blocks are known. Therefore, the size of the search space is $\binom{\binom{v}{k}}{b}$. It explains immediately the difficulty of finding designs by explicit constructions. To reduce the search space, additional constraints might be taken into account. If an automorphism group acts on a design, it acts on the set of points X and on the set of blocks $\mathcal{B}$, partitioning them into orbits. It acts on the set $\binom{X}{k}$ from which we choose our blocks as well, partitioning it into orbits of k-subsets. Clearly, only complete orbits of blocks can be chosen to keep the automorphism group acting on the design. That fact reduces the search space size enormously.

The chosen orbits form a tactical decomposition of the incidence matrix. Out of that fact one can determine

⋆ This work has been supported by the Croatian Science Foundation under the projects 6732 and 9752.

additional necessary conditions for the existence of such a decomposition and reduce the search space further (for details, see Krčadinac et al. (2014)). The most general and successfully widely used method for constructing t-designs by computer goes back to Kramer et al. (1976), where the problem is set as an integer system of linear equations.

3. MOSAICS OF DESIGNS

The rather unsatisfactory fact that a systematic approach towards constructing and classifying designs is often not possible due to the size of the search space, we came to the idea of combining designs together. Looking at Example 1, if we treat the 0's as incidences, we can interprete the $v \times b$ incidence matrix as filled with incidences of two (here complementary) designs. There is place for generalization.

Example 2. Here the matrix of order 7 is filled with 3 different numbers

$$M = \begin{bmatrix} 0&1&1&2&1&2&2\\ 2&0&1&1&2&1&2\\ 2&2&0&1&1&2&1\\ 1&2&2&0&1&1&2\\ 2&1&2&2&0&1&1\\ 1&2&1&2&2&0&1\\ 1&1&2&1&2&2&0 \end{bmatrix}$$

and if we interpret them as incidences, we can clearly write this decomposition as

$$2\text{-}(7,1,0) \oplus 2\text{-}(7,3,1) \oplus 2\text{-}(7,3,1).$$

It is namely obvious that this matrix "consists" of the following three incidence matrices

$$M_0 = \begin{bmatrix} 1&0&0&0&0&0&0\\ 0&1&0&0&0&0&0\\ 0&0&1&0&0&0&0\\ 0&0&0&1&0&0&0\\ 0&0&0&0&1&0&0\\ 0&0&0&0&0&1&0\\ 0&0&0&0&0&0&1 \end{bmatrix}, \quad M_1 = \begin{bmatrix} 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1\\ 1&1&0&1&0&0&0 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} 0&0&0&1&0&1&1\\ 1&0&0&0&1&0&1\\ 1&1&0&0&0&1&0\\ 0&1&1&0&0&0&1\\ 1&0&1&1&0&0&0\\ 0&1&0&1&1&0&0\\ 0&0&1&0&1&1&0 \end{bmatrix},$$

where $M_0 + M_1 + M_2 = J$, J, being the all-one matrix.

In Gnilke et al. (2018) we have introduced the following combinatorial object. Let c be a positive integer and let $\mathcal{B}_i$ be designs with parameters t_i-(v, k_i, λ_i), $i = 1, \ldots, c$ with the same number of points v and blocks b. A *c-mosaic of designs* $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_c$ is a $(v \times b)$ matrix $M = [m_{pq}]$, $m_{pq} \in \{l_1, l_2, \ldots, l_c\}$ for which holds that matrices M_i defined as

$$[M_i]_{pq} = \begin{cases} 1, & m_{pq} = l_i \\ 0, & otherwise \end{cases}$$

are incidence matrices of designs $\mathcal{B}_i$.

Once having defined mosaics of designs as a tiling of a $(v \times b)$ matrix, the next task is again to find explicit constructions of these objects.

A design $\mathcal{B}$ on a set X is called *resolvable* if there exists a partition of the set of blocks $\mathcal{B}$ into so called parallel classes, such that every class itself is a partition of the set of points X. Note that for a resolvable t-(v, k, λ) design the number of parallel classes is $p = \lambda \binom{v-1}{t-1} \binom{k-1}{t-1}^{-1}$ and each class contains $\frac{v}{k}$ blocks.

Theorem 1. Let D be the incidence matrix of a resolvable t-(v, k, λ) design, where the columns have been arranged by parallel classes. Let L be a latin square of order $\frac{v}{k}$ with entries $l_1, \ldots, l_{\frac{v}{k}}$. Then $M := D(I_p \otimes L)$ is a $\frac{v}{k}$-mosaic.

Corollary 2. Let F be the field with q elements. Then there is a q-mosaic of affine planes of order q:

$$2\text{-}(q^2, q^2, q^2 + q) = 2\text{-}(q^2, q, 1) \oplus \cdots \oplus 2\text{-}(q^2, q, 1).$$

Open question. Is there a mosaic of a matrix of order 31 of the following form (with the following parameters):

$$2\text{-}(31, 15, 7) \ \oplus \ 2\text{-}(31, 10, 3) \ \oplus \ 2\text{-}(31, 6, 1)?$$

We are still unable to answer this question because of the number of combinatorial possibilities and at the same time couldn't find an argument against the existence.

Nice application of mosaics of designs have been described in Wiese et al. (2022).

4. DESIGN CUBES

Another idea how to combine designs together is looking at 3-dimensional incidence matrices $A = [a_{ijk}]$, $i, j, k = 1, \ldots, v$, $a_{ijk} \in \{0, 1\}$. We will not introduce a third kind of objects (although this might be possible and find its applications), but we want for each 2-dimensional submatrix to be incidence matrix of a design. Hence, for a given A, take a fixed i (from 1 to v), and observe the $v \times v$ matrix $A^i_{jk} = [a_{ijk}]$, $j, k = 1, \ldots, v$ which needs to be an incidence matrix of a (symmetric) design. Do the same for a fixed $j = 1, \ldots, v$ and fixed $k = 1, \ldots, v$. If all $3v$ matrices A^i_{jk}, A^j_{ik} and A^k_{ij} are incidence matrices of a design, we shall call such a 3-dimensional matrix A a *design cube*. If the parameters of the symmetric design are 2-(v, k, λ), we shall assign this parameter triple to the design cube as well.

Example 3. We continue with combining the incidence matrix of the Fano plane from Example 1 in order to get a design cube. The first matrix below is just a copy of it, and the six others are row-shifts of that matrix.

$$\begin{bmatrix} 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1\\ 1&1&0&1&0&0&0 \end{bmatrix} \begin{bmatrix} 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1\\ 1&1&0&1&0&0&0\\ 0&1&1&0&1&0&0 \end{bmatrix} \begin{bmatrix} 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1\\ 1&1&0&1&0&0&0\\ 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0 \end{bmatrix}$$

$$\begin{bmatrix} 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1\\ 1&1&0&1&0&0&0\\ 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1 \end{bmatrix} \begin{bmatrix} 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1\\ 1&1&0&1&0&0&0\\ 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0 \end{bmatrix} \begin{bmatrix} 1&0&1&0&0&0&1\\ 1&1&0&1&0&0&0\\ 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1 \end{bmatrix}$$

$$\begin{bmatrix} 1&1&0&1&0&0&0\\ 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1 \end{bmatrix}$$

Imagine putting these matrices one above the other, playing the role of $A^1_{ij}, \ldots, A^7_{ij}$ (the third index fixed) from the definition above. We get a cube of order 7. We need to check that all 2-dimensional (sub)matrices of order 7 are incidence matrices of a 2-$(7, 3, 1)$ design. This is trivial if we fix the 3. dimension "k": all the matrices listed above are incidence matrices of the Fano plane. If we fix the first dimension (index) i, we see that the i-th rows are simply shifts of the previous row. Finally, if we fix the second dimension (index) j, we see that columns are shifts as well, but in the opposite direction.

Theorem 3. If there exists a (v, k, λ) difference set, then there exists a design cube with parameters 2-(v, k, λ).

The proof of this theorem follows facts listed in Example 3 and can be generalized easily. Namely, the incidence matrix of the Fano plane in our example is presented as a cyclic matrix hence it can be seen as the *development* of a $(7, 3, 1)$ *difference set* (Colbourn et al. (2007)) in the cyclic group of order 7.

For the moment we are not aware of any examples not coming from difference sets and finding such an example would be the first task in continuing the research on design cubes. Furthermore, it would be interesting to define and find generalizations of design cubes in higher dimensions.

REFERENCES

Colbourn, C. J., Dinitz, J. H. (2007). *Handbook of Combinatorial Designs (2nd ed.)*. Chapman & Hall / CRC, Boca Raton.

Gnilke, O. W., Greferath, M., and Pavčević, M. O. (2018). Mosaics of combinatorial designs. *Designs, codes and cryptography*, 86, 85–95.

Kramer, E. S. and Mesner, D. M. (1976). t-designs on hypergraphs. *Discrete Mathematics*, 15, 263–296.

Krčadinac, V., Nakić, A., Pavčević, M. O. (2014). Equations for coefficients of tactical decomposition matrices for t-designs. *Designs, codes and cryptography*, 72, 465–469.

Wiese, M., and Boche, H. (2022). Mosaics of combinatorial designs for information-theoretic security. *Designs, codes and cryptography*, online.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Robust Fundamental Lemma for Data-driven Control

Jeremy Coulson * **Henk van Waarde** ** **Florian Dörfler** *

* *Automatic Control Laboratory, ETH Zürich, ZH 8006, Switzerland (e-mail: {jcoulson, dorfler}@control.ee.ethz.ch).*
** *Systems, Control and Optimization group, University of Groningen, 9747 AG Groningen, The Netherlands (e-mail: h.j.van.waarde@rug.nl).*

Abstract: The fundamental lemma by Willems and coauthors facilitates a parameterization of all trajectories of a linear time-invariant system in terms of a single, measured one. This result plays an important role in data-driven simulation and control. Under the hood, the fundamental lemma works by applying a persistently exciting input to the system. This ensures that the Hankel matrix of resulting input/output data has the "right" rank, meaning that its columns span the entire subspace of trajectories. However, such binary rank conditions are known to be fragile in the sense that a small additive noise could already cause the Hankel matrix to have full rank. Therefore, in this extended abstract we present a robust version of the fundamental lemma. The idea behind the approach is to guarantee certain lower bounds on the singular values of the data Hankel matrix, rather than mere rank conditions. This is achieved by designing the inputs of the experiment such that the minimum singular value of a deeper input Hankel matrix is sufficiently large. This inspires a new quantitative and robust notion of persistency of excitation. The relevance of the result for data-driven control will also be highlighted through comparing the predictive control performance for varying degrees of persistently exciting data.

Keywords: Identification, linear systems, data-driven control

1. INTRODUCTION

The fundamental lemma from Willems et al. (2005) is a powerful result that enables the characterization of the subspace of all possible trajectories of a linear time-invariant (LTI) system using raw time series data sorted into a Hankel matrix. The result has inspired many methods for data-driven analysis and control; see (Markovsky and Rapisarda, 2008; van Waarde et al., 2020; van Waarde, 2021; Coulson et al., 2019; De Persis and Tesi, 2019; Berberich et al., 2020), and the survey by Markovsky and Dörfler (2021) and references therein.

At its core, the fundamental lemma requires applying an input sequence to the system that is persistently exciting of sufficient order such that the resulting input/output data Hankel matrix spans the entire subspace of possible trajectories of the system. In other words, if a deeper input data Hankel matrix is full row rank, the resulting input/output data Hankel matrix spans the admissible trajectory subspace. This input/output data Hankel matrix can then be used as a non-parametric system model.

One major drawback of the fundamental lemma is that it only holds for noise-free data. Indeed, when the data are corrupted by noise, rank conditions are no longer sufficient for the data matrix to span the admissible trajectory subspace leading to poor performance when used for data-driven analysis and control. This demonstrates that these binary rank conditions can no longer indicate suitable data when in the presence of noise. This motivates defining a new *quantitative* notion of persistency of excitation that gives a measure of *how* persistently exciting the inputs are. In adaptive control, such quantitative notions are studied (Åström and Wittenmark, 2008, Remark 1, pg. 64), but not in the context of the fundamental lemma.

In this extended abstract we propose a robust fundamental lemma that relies on a new *quantitative* notion of persistency of excitation. The result informs the input selection such that the minimum singular value of the input/output data matrix is lower bounded by a user specified parameter. This results in a data matrix that is more robust to noise leading to better performance when used for data-driven analysis and control.

The rest of the extended abstract is organized as follows. We begin with notation. Section 2 formalizes the problem of interest. Section 3 contains the main result whose relevance is illustrated through a data-driven control case study in Section 4. We conclude in Section 5.

Notation: Let $m, n \in \mathbb{Z}_{>0}$. Given a matrix $M \in \mathbb{R}^{m\times n}$ and integer $i \in \mathbb{Z}_{>0}$ we denote the i-th singular value of M by $\sigma_i(M)$ with the ordering $0 \leq \sigma_1(M) \leq \sigma_2(M) \leq \cdots \leq \sigma_{\min\{m,n\}}$. When $m = n$ and $M = M^\top$, we denote the i-th eigenvalue of M by $\lambda_i(M)$ with the ordering $\lambda_1(M) \leq \lambda_2(M) \leq \cdots \lambda_n(M)$. We use $\|M\| = \sigma_{\min\{m,n\}}$ to denote the spectral norm of a matrix M, $\|M\|_F$ the Frobenius norm, and $\|x\|$ to denote the 2-norm of a vector

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

x. The Moore-Penrose pseudo inverse of M is denoted by $M^\dagger$. Given $i, j, T \in \mathbb{Z}_{\geq 0}$ with $i \leq j$ and a sequence $\{z(t)\}_{t=0}^{T-1} \subset \mathbb{R}^n$ define

$$z_{[i,j]} := [z(i)^\top \; z(i+1)^\top \cdots z(j)^\top]^\top.$$

Given an integer $k \in \mathbb{Z}_{>0}$ with $k \leq j - i + 1$, define the *Hankel matrix of depth* k associated with $z_{[i,j]}$ as

$$\mathcal{H}_k(z_{[i,j]}) := \begin{bmatrix} z(i) & z(i+1) & \cdots & z(j-k+1) \\ z(i+1) & z(i+2) & \cdots & z(j-k+2) \\ \vdots & \vdots & \ddots & \vdots \\ z(i+k-1) & z(i+k) & \cdots & z(j) \end{bmatrix}.$$

2. PROBLEM STATEMENT

Consider the discrete-time LTI system

$$x(t+1) = Ax(t) + Bu(t), \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state, and $u(t) \in \mathbb{R}^m$ is the control input. Throughout this extended abstract we will focus on a special case of full state measurements. We recall the definition of persistency of excitation.

Definition 2.1. Let $T \in \mathbb{Z}_{>0}$. The input sequence $u_{[0,T-1]}$ is called persistently exciting of order $k \in \mathbb{Z}_{>0}$ if $\mathcal{H}_k(u_{[0,T-1]})$ has full row rank.

We now state a version of the fundamental lemma (Willems et al., 2005) for the special case of input/state data.

Theorem 2.1. Let (A, B) be controllable and $T \in \mathbb{Z}_{>0}$. Let $(u_{[0,T-1]}, x_{[0,T-1]})$ be an input/state trajectory of (1) such that $u_{[0,T-1]}$ is persistently exciting of order $n+1$. Then the data matrix

$$\begin{bmatrix} \mathcal{H}_1(x_{[0,T-1]}) \\ \mathcal{H}_1(u_{[0,T-1]}) \end{bmatrix} = \begin{bmatrix} x(0) & x(1) & \cdots & x(T-1) \\ u(0) & u(1) & \cdots & u(T-1) \end{bmatrix} \tag{2}$$

has rank $n + m$.

As a result, the matrix (2) along with $x(T)$ fully capture the behaviour of (1) and can be used for data-driven analysis and control or system identification (Markovsky and Dörfler, 2021). However, the above rank condition on the data matrix (2) is not always a valid indicator that the data is suitable for characterizing the behaviour of (1). In fact, when the data is corrupted by noise, (2) may have full rank, but can lead to poor performance when used for data-driven analysis or control. We present a motivating example outlining precisely how we can improve on such rank conditions.

Example 2.1. Suppose we wish to identify matrices A, B of the system

$$x(t+1) = Ax(t) + Bu(t) + w(t), \tag{3}$$

where $w(t) \in \mathbb{R}^n$ is a noise term. Let $(u_{[0,T-1]}, x_{[0,T-1]})$ be an input/state trajectory of the noisy system such that $u_{[0,T-1]}$ is persistently exciting of order $n+1$. To identify A, B, we consider the least squares problem

$$\min_{A,B} \quad \left\| \mathcal{H}_1(x_{[1,T]}) - [A \; B] \begin{bmatrix} \mathcal{H}_1(x_{[0,T-1]}) \\ \mathcal{H}_1(u_{[0,T-1]}) \end{bmatrix} \right\|_F$$

with solution given by

$$\begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} := \mathcal{H}_1(x_{[1,T]}) \begin{bmatrix} \mathcal{H}_1(x_{[0,T-1]}) \\ \mathcal{H}_1(u_{[0,T-1]}) \end{bmatrix}^\dagger. \tag{4}$$

The data satisfies

$$\mathcal{H}_1(x_{[1,T]}) = [A \; B] \begin{bmatrix} \mathcal{H}_1(x_{[0,T-1]}) \\ \mathcal{H}_1(u_{[0,T-1]}) \end{bmatrix} + \mathcal{H}_1(w_{[0,T-1]}).$$

Hence, the error of our estimate is given by

$$\begin{aligned} \left\| \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} - [A \; B] \right\| &= \left\| \mathcal{H}_1(w_{[0,T-1]}) \begin{bmatrix} \mathcal{H}_1(x_{[0,T-1]}) \\ \mathcal{H}_1(u_{[0,T-1]}) \end{bmatrix}^\dagger \right\| \\ &\leq \frac{\sigma_n(\mathcal{H}_1(w_{[0,T-1]})}{\sigma_1\left(\begin{bmatrix} \mathcal{H}_1(x_{[0,T-1]}) \\ \mathcal{H}_1(u_{[0,T-1]}) \end{bmatrix} \right)}. \end{aligned}$$

Note that the estimation error can be arbitrarily large, regardless of whether (2) is full row rank. What is important instead is the smallest singular value of (2). Indeed, if we wish to identify A, B to within some specified error, we require that the smallest singular value of the data matrix (2) is bounded below by some threshold depending on the noise term w. •

This example motivates us to develop a robust fundamental lemma that moves away from rank conditions by defining a quantitative notion of persistency of excitation which guarantees a lower bound on the minimum singular value of the data matrix (2). More formally, the goal of this extended abstract is to solve the following problem.

Problem 2.1. Let $\delta > 0$. Design an input sequence $u_{[0,T-1]}$ such that the resulting data matrix satisfies

$$\sigma_1\left(\begin{bmatrix} \mathcal{H}_1(x_{[0,T-1]}) \\ \mathcal{H}_1(u_{[0,T-1]}) \end{bmatrix} \right) \geq \delta. \tag{5}$$

3. MAIN RESULT

The two main ingredients of the fundamental lemma are controllability and persistency of excitation. To establish a robust fundamental lemma, we must develop quantitative notions of these two main ingredients. We start with persistency of excitation.

Definition 3.1. Let $T \in \mathbb{Z}_{>0}$, $\alpha > 0$. The input sequence $u_{[0,T-1]}$ is called α-persistently exciting of order $k \in \mathbb{Z}_{>0}$ if $\sigma_1(\mathcal{H}_k(u_{[0,T-1]})) \geq \alpha$.

This is a natural generalization of persistency of excitation since for any $\alpha > 0$ an α-persistently exciting signal of order k is necessarily persistently exciting of order k.

We now focus on the second main ingredient of the fundamental lemma: controllability. Define matrix $M \in \mathbb{R}^{(n+m+nm)\times(n+m+nm)}$ and vector $z \in \mathbb{R}^{n+m+nm}$ as

$$M := \begin{bmatrix} A & B & 0 & \cdots & 0 \\ 0 & 0 & I_m & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I_m \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad z = \begin{bmatrix} \xi \\ \eta \\ 0_{nm} \end{bmatrix} \tag{6}$$

where $\xi \in \mathbb{R}^n$, $\eta \in \mathbb{R}^m$, and I_m and 0_{nm} denote the $m \times m$ identity matrix and the zero vector of length nm, respectively. Define for any z of the form (6)

$$\Theta_z := \begin{bmatrix} z & M^\top z & \cdots & (M^\top)^n z \end{bmatrix}^\top. \tag{7}$$

Note that $\Theta_z^\top$ can be viewed as an extended controllability matrix of the pair $(M^\top, z)$. We make the following quantitative controllability assumption on this pair.

Assumption 3.1. Let $\rho > 0$. For all z of the form (6) with $\|z\| = 1$, assume

$$\sigma_1(\Theta_z) \geq \rho. \tag{8}$$

This assumption can be thought of as ensuring that the finite horizon controllability Gramian $\Theta_z^\top \Theta_z = \sum_{j=0}^{n} (M^\top)^j z z^\top M^j$ has lower bounded eigenvalues. This assumption is not restrictive since for any controllable (A, B), there exists uniform $\rho_0 > 0$ that lower bounds $\sigma_1(\Theta_z)$ for all z with $\|z\| = 1$, by the following lemma.

Lemma 3.1. Let (A, B) be controllable. Then

(i) $\operatorname{rank}(\Theta_z) = n + 1$, and
(ii) there exists $\rho_0 > 0$ such that $\sigma_1(\Theta_z) \geq \rho_0$,

for all z be of the form (6) with $\|z\| = 1$.

Proof. We begin by proving (i). Assume on the contrary that the rows of Θ_z are linearly dependent. From the structure of Θ_z, we must have that $\eta = 0$ and thus $z = (\xi, 0_{(n+1)m})$. Likewise, we must have that $\xi^\top B = \xi^\top AB = \cdots = \xi^\top A^{n-1} B = 0$. By controllability of (A, B), this implies that $\xi = 0$, and hence $z = 0$. This contradicts the fact that $\|z\| = 1$, proving the claim. We now prove (ii). Assume on the contrary that $\forall \rho_0 > 0$, $\exists\ z$ with $\|z\| = 1$ such that $\sigma_1(\Theta_z) < \rho_0$. Let $\{\rho_j\}_{j=1}^{\infty} \subset \mathbb{R}$ be a sequence such that $\lim_{j\to\infty} \rho_j = 0$. By assumption, for each ρ_j, there exists z_j with $\|z_j\| = 1$ and $\sigma_1(\Theta_{z_j}) < \rho_j$. Since the set $\{z \mid \|z\| = 1\}$ is compact, $\{z_j\}_{j=1}^{\infty}$ has a convergent subsequence converging to some $\bar{z}$ with $\|\bar{z}\| = 1$. Thus, $\sigma_1(\Theta_{\bar{z}}) \leq 0$. However, by part (i), $\Theta_{\bar{z}}$ has full row rank and hence, there exists $\bar{\rho} > 0$ such that $\sigma_1(\Theta_{\bar{z}}) > \bar{\rho}$ which is a contradiction, proving the result. ■

We now state the main theorem which solves Problem 2.1.

Theorem 3.1. Let $T \in \mathbb{Z}_{>0}$, (A, B) be controllable, and $\delta > 0$. Suppose that Assumption 3.1 holds and $(u_{[0,T-1]}, x_{[0,T-1]})$ is an input/state trajectory of (1) such that $u_{[0,T-1]}$ is $\delta\frac{\sqrt{n+1}}{\rho}$-persistently exciting of order $n+1$. Then

$$\sigma_1\left(\begin{bmatrix}\mathcal{H}_1(x_{[0,T-1]})\\ \mathcal{H}_1(u_{[0,T-1]})\end{bmatrix}\right) \geq \delta.$$

Proof. Denote the minimum singular value of the input/state data matrix (2) by $\sigma \geq 0$ with corresponding left and right singular vectors $(\xi, \eta) \in \mathbb{R}^{n+m}$ and $v \in \mathbb{R}^T$, respectively. Then, $\|(\xi, \eta)\| = \|v\| = 1$ and

$$[\xi^\top\ \eta^\top] \begin{bmatrix}\mathcal{H}_1(x_{[0,T-1]})\\ \mathcal{H}_1(u_{[0,T-1]})\end{bmatrix} = \sigma v^\top.$$

Let $z = (\xi, \eta, 0_{nm})$. By definition of Θ_z in (9) and the dynamics (1), we have

$$\Theta_z \begin{bmatrix}\mathcal{H}_1(x_{[0,T-n-1]})\\ \mathcal{H}_{n+1}(u_{[0,T-1]})\end{bmatrix} = \sigma \mathcal{H}_{n+1}(v_{[0,T-1]}). \tag{9}$$

By Cauchy's interlacing theorem (Horn and Johnson, 1994, Corollary 3.1.3),

$$\sigma_1\left(\mathcal{H}_{n+1}(u_{[0,T-1]})\right) \leq \sigma_{n+1}\left(\begin{bmatrix}\mathcal{H}_1(x_{[0,T-n-1]})\\ \mathcal{H}_{n+1}(u_{[0,T-1]})\end{bmatrix}\right). \tag{10}$$

By the Courant-Fischer-Weyl max-min principle (Horn and Johnson, 1994, Theorem 3.1.2),

$$\sigma_{n+1}\left(\begin{bmatrix}\mathcal{H}_1(x_{[0,T-n-1]})\\ \mathcal{H}_{n+1}(u_{[0,T-1]})\end{bmatrix}\right) = \min_{\substack{\mathcal{U}\\ \dim(\mathcal{U})=n+1}} \max_{\substack{y\in\mathcal{U}\\ \|y\|=1}} \left\| \begin{bmatrix}\mathcal{H}_1(x_{[0,T-n-1]})\\ \mathcal{H}_{n+1}(u_{[0,T-1]})\end{bmatrix}^\top y \right\|.$$

By Lemma 3.1, $\operatorname{rank}(\Theta_z) = n + 1$, and hence

$$\begin{aligned}
&\sigma_{n+1}\left(\begin{bmatrix}\mathcal{H}_1(x_{[0,T-n-1]})\\ \mathcal{H}_{n+1}(u_{[0,T-1]})\end{bmatrix}\right)\\
&\leq \max_{\substack{y\in\operatorname{im}\Theta_z^\top\\ \|y\|=1}} \left\| \begin{bmatrix}\mathcal{H}_1(x_{[0,T-n-1]})\\ \mathcal{H}_{n+1}(u_{[0,T-1]})\end{bmatrix}^\top y \right\|\\
&= \max_{\|\Theta_z^\top q\|=1} \left\|\sigma \mathcal{H}_{n+1}(v_{[0,T-1]})^\top q\right\|\\
&\leq \sigma \left\|\mathcal{H}_{n+1}(v_{[0,T-1]})^\top\right\| \max_{\|\Theta_z^\top q\|=1} \|q\|\\
&\leq \sigma\sqrt{n+1} \max_{\|\Theta_z^\top q\|=1} \|q\|,
\end{aligned} \tag{11}$$

where the equality comes from (9) and the last inequality holds because the rows of $\mathcal{H}_{n+1}(v_{[0,T-1]})$ have norm at most 1. Without loss of generality, write $q = \sum_{j=1}^{n+1} \alpha_j \mu_j$ where $\alpha_j \in \mathbb{R}$ and $\mu_j \in \mathbb{R}^{n+1}$ are orthonormal eigenvectors of $\Theta_z \Theta_z^\top$ corresponding to eigenvalues $\lambda_j(\Theta_z \Theta_z^\top)$. Then

$$\max_{\|\Theta_z^\top q\|=1} \|q\|^2 = \max_{\sum_{j=1}^{n+1} \alpha_j^2 \lambda_j(\Theta_z\Theta_z^\top)=1} \sum_{j=1}^{n+1} \alpha_j^2.$$

We see the maximum is achieved for $\alpha_1 = \pm\frac{1}{\sqrt{\lambda_1(\Theta_z\Theta_z^\top)}}$, and $\alpha_j = 0$ for $j \in \{2, \ldots, n+1\}$. Hence,

$$\max_{\|\Theta_z^\top q\|=1} \|q\| = \frac{1}{\sqrt{\lambda_1(\Theta_z\Theta_z^\top)}}.$$

Combining the above with (10) and (11), we obtain

$$\sigma \geq \sigma_1\left(\mathcal{H}_{n+1}(u_{[0,T-1]})\right) \sqrt{\frac{\lambda_1(\Theta_z\Theta_z^\top)}{n+1}}.$$

Substituting $\sqrt{\lambda_1(\Theta_z\Theta_z^\top)} = \sigma_1(\Theta_z) \geq \rho$ by Assumption 3.1 and using the fact that $u_{[0,T-1]}$ is $\delta\frac{\sqrt{n+1}}{\rho}$-persistently exciting of order $n+1$ yields the result. ■

The theorem tells us how the inputs should be chosen such that *for any* user defined parameter δ, the smallest singular value of the data matrix (2) is lower bounded by δ. The degree of persistency of excitation needed depends on ρ which is assumed to be a prior in our setting. However, Lemma 3.1 shows that, for any controllable system (1), there exists $\rho_0 > 0$ such that $\sigma_1(\Theta_z) \geq \rho_0$ for all z. To design an input sequence so that (5) holds, we only require a lower bound on ρ_0 since any input that is $\delta\frac{\sqrt{n+1}}{\rho_0}$-persistently exciting of order $n+1$ yields the desired result.

4. NUMERICAL EXAMPLE

In this section we compare the performance of several data sets with varying degrees of persistency of excitation for a data-driven control task. Our hypothesis is that data sets whose inputs have a larger degree of persistency excitation will perform better when the data is corrupted by noise. Consider a controllable system (3) with

$$A = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix},\ B = \begin{bmatrix}0\\ 1\end{bmatrix}.$$

We generated 3 input data sequences of length $T = 50$ with varying degrees of persistency of excitation. We denote

the i-th input data sequence by $u^{(i)}_{[0,T-1]}$. The inputs for each input data sequence were chosen as $u^{(1)}(t) \sim \mathcal{N}(0, 10^{-2}), u^{(2)}(t) = 0.05u^{(1)}(t), u^{(3)}(t) = 0.01u^{(1)}(t)$ for all $t \in \{0, \dots, T-1\}$. As a result, we obtained 3 input sequences that were $\alpha^{(i)}$-persistently exciting of order $n+1$, with $\alpha^{(1)} = 0.48$, $\alpha^{(2)} = 0.024$, $\alpha^{(3)} = 0.0048$. The corresponding state data $x^{(i)}_{[0,T]}$ was generated by (3) where the noise $w(t) \sim \mathcal{N}(0, 10^{-4}I_n)$ was the same across all data sets. Using the data, we constructed 3 different data-driven one-step predictors as in (4)

$$x(t+1) = \mathcal{H}_1\left(x^{(i)}_{[1,T]}\right) \begin{bmatrix} \mathcal{H}_1\left(x^{(i)}_{[0,T-1]}\right) \\ \mathcal{H}_1\left(u^{(i)}_{[0,T-1]}\right) \end{bmatrix}^{\dagger} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}. \quad (12)$$

We then used the predictor equations (12) for a predictive control reference tracking task by solving the following optimization problem in a receding horizon fashion:

$$\begin{aligned} \min_{x,u} \quad & \sum_{k=0}^{T_\mathrm{f}-1} \|x_k - r\|^2 + \|u_k\|^2 \\ \text{s.t.} \quad & x_0 = x(t) \\ & x_{k+1} = \mathcal{H}_1\left(x^{(i)}_{[1,T]}\right) \begin{bmatrix} \mathcal{H}_1\left(x^{(i)}_{[0,T-1]}\right) \\ \mathcal{H}_1\left(u^{(i)}_{[0,T-1]}\right) \end{bmatrix}^{\dagger} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \\ & k \in \{0, \dots, T_\mathrm{f} - 1\} \end{aligned} \quad (13)$$

where $x(t)$ denotes the current state at time t, $r \in \mathbb{R}^n$ is the reference, and $T_\mathrm{f} = 10$ is the prediction horizon. System (3) was simulated for each data set with noise $w(t) \sim \mathcal{N}(0, 10^{-4}I)$ the same across all simulations and inputs obtained by solving (13) in receding horizon. Figure 1 depicts the performance of the three data-driven one-step predictors for the reference tracking task.

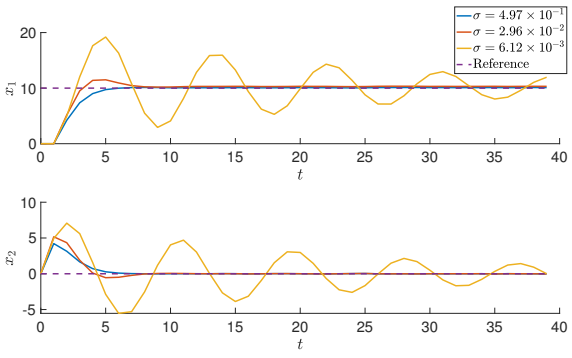


Fig. 1. State trajectories $x(t) = (x_1(t), x_2(t))$ controlled by solving (13) in a receding horizon fashion for three data sets with varying degrees of persistency of excitation. The legend indicates the minimum singular value of the data matrices (2) used for prediction in (13).

For this example, we approximated ρ_0 for which Lemma 3.1 holds. By randomly sampling vectors z with $\|z\| = 1$, we obtained $\rho_0 \approx 0.105$. Thus, Assumption 3.1 should hold for any $\rho \leq \rho_0$. Based on ρ_0 and the degree of persistency of excitation of each input sequence, we can compute the value $\delta^{(i)}$ for which (5) holds for the i-th input/state data. Theorem 3.1 guarantees that (5) holds for each data set with $\delta^{(1)} = 2.9 \times 10^{-2}$, $\delta^{(2)} = 1.4 \times 10^{-3}$, and $\delta^{(3)} = 2.9 \times 10^{-4}$. As we see from Figure 1, the performance of the data-driven predictors decreases as the minimum singular value of the data matrix (2) decreases. This can be attributed to the fact that the one-step predictor in (12) becomes more accurate as the smallest singular value of the data matrix decreases (as seen in Example 2.1), leading to better performance.

5. CONCLUSION

In this extended abstract, we defined a new quantitative notion of persistency of excitation by analyzing the minimum singular value of an input Hankel matrix. Based on this notion, we specified to what degree the inputs should be persistently exciting to ensure that the smallest singular value of the resulting input/state matrix is larger than a user defined threshold, thus generalizing the celebrated fundamental lemma. As a result, we are able to move away from classical rank conditions and give a quantitative notion of data suitability. A comparison of the control performance of several data sets with varying degrees of persistently exciting input data suggested that data being generated by inputs with a larger degree of persistency of excitation are better suited to control tasks. Future work includes extending these results to the general input/output case.

REFERENCES

Åström, K.J. and Wittenmark, B. (2008). *Adaptive control*. Dover Publications, 2nd edition.

Berberich, J., Köhler, J., Müller, M.A., and Allgöwer, F. (2020). Data-driven model predictive control with stability and robustness guarantees. *IEEE Transactions on Automatic Control*, 66(4), 1702–1717.

Coulson, J., Lygeros, J., and Dörfler, F. (2019). Data-enabled predictive control: In the shallows of the DeePC. In *2019 18th European Control Conference (ECC)*, 307–312.

De Persis, C. and Tesi, P. (2019). Formulas for data-driven control: Stabilization, optimality and robustness. *IEEE Transactions on Automatic Control*.

Horn, R.A. and Johnson, C.R. (1994). *Topics in matrix analysis*. Cambridge university press.

Markovsky, I. and Dörfler, F. (2021). Behavioral systems theory in data-driven analysis, signal processing, and control. *Annual Reviews in Control*, 52, 42–64.

Markovsky, I. and Rapisarda, P. (2008). Data-driven simulation and control. *International Journal of Control*, 81(12), 1946–1959.

van Waarde, H.J. (2021). Beyond persistent excitation: Online experiment design for data-driven modeling and control. *IEEE Control Systems Letters*.

van Waarde, H.J., De Persis, C., Camlibel, M.K., and Tesi, P. (2020). Willems' fundamental lemma for state-space systems and its extension to multiple datasets. *IEEE Control Systems Letters*, 4(3), 602–607.

Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B.L. (2005). A note on persistency of excitation. *Systems & Control Letters*, 54(4), 325–329.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Aggregating distributed energy resources for grid flexibility services: a distributed game theoretic approach

Xiupeng Chen* **Jacquelien M. A. Scherpen*,**** **Nima Monshizadeh***

**Engineering and Technology institute Groningen, University of Groningen, 9747 AG Groningen, The Netherlands*
***Jan C. Willems Center for Systems and Control, University of Groningen, 9747 AG Groningen, Groningen, the Netherlands*
e-mail: {xiupeng.chen, j.m.a.scherpen, n.monshizadeh}@rug.nl.

Abstract: We propose a novel fully decentralized energy management scheme for aggregating distributed energy resources for grid flexibility services in wholesale electricity market. We model this problem as a multi-leader-multi-follower noncooperative game. Then a fully distributed algorithm in discrete-time is proposed to solve the problem and find the Nash Equilibrium(NE). In this algorithm, each aggregator only needs to exchange its estimate of the aggregate and an auxiliary variable with its neighbours. This scheme shows the scalability and efficiency in aggregating flexibility services from a large number of prosumers.

Keywords: Game theory, Smart grid, Flexibility service, Nash equilibrium seeking.

1. INTRODUCTION

In recent years, the growing climate and environmental concerns have led to a rapid increase of the contribution of renewable energy capacities in electricity generation. However, the variability and uncertainty of renewable generation bring potential challenges in the operation of power systems and influence the performance and outcome of electricity market. Designing energy management mechanisms for Distributed Energy Resources(DERs) in the smart grid has been considered as a potential solution for a renewable-power future. The smart grid is typically composed of a variety of new participants, such as microgrids, aggregators, and prosumers. Each of them is self-interested and has different capabilities and objectives. This heterogeneous nature of the smart grid motivates the adoption of game theory as an analytical tool to study the interactions among them.

In particular, managing DERs to response to the system's overall condition is a powerful solution to increasing grid flexibility and facilitating the integration of renewable generations. This will enable end-consumers to take up an unprecedented proactive role(i.e. prosumers) and reap financial benefits through leveraging their flexibility by providing services to system operators. However, it is difficult to imagine that all these individual prosumers directly participate in the wholesale electricity market, thus a market participant, the aggregator, has been introduced to manage these DERs locally and ensure no distribution constraints are violated (Gkatzikis et al. (2013)). There are two main approaches which an aggregator can employ to steer the prosumers to the optimal operation point. One is the fully centralized scheme, where the aggregator has access to all parameters of the prosumers (Parvania et al. (2013)). The other one is pricing based methods, where the aggregator is treated as a leader and proposes some prices to the following prosumers (see e.g. Zugno et al. (2013)).

Competition among participating agents can be incoporated in a game theoretic setup, where the aggregator aims to steer the prosumers to a desired setpoint, often taken to be the Nash equilibrium of the game. Designing algorithms for equilibrium-seeking problems has attracted substantial research interest in the last decades. These algorithms can be roughly divided into two categories in terms of methods used: gradient-based algorithms (Franci and Grammatico (2020), Pavel (2019)) and proximal-point algorithms (Belgioioso and Grammatico (2019), Yi and Pavel (2018)). In these works, each player is required to know or estimate the overall information of all the other players. The algorithm in (Koshal et al. (2016)) only requires each player maintains an estimate of the true aggregate, but it requires diminishing step-sizes for exact convergence. The work (Gadjov and Pavel (2020)) proposes a single-layer distributed algorithm that reaches a variational generalized NE under constant step sizes while the initial values of aggregate estimations must be same with those of actions. In (De Persis and Grammatico (2019), Shakarami et al. (2019)), continuous-time algorithms for aggregative games are proposed. While discrete-time algorithms are in general more suited for implementations in energy-domain applications, their design requires additional care, e.g. a careful choice of step sizes is needed.

This work consists of two parts: network modeling and game-theoretic algorithms. In the first part, we consider a fully decentralized energy management scheme for aggregating distributed energy resources for grid flexibility services in wholesale electricity market. We explicitly model

this scheme as a multi-leader-multi-follower game. In the upper level, the aggregators aim to maximize their profits by incentivizing the prosumers to utilize their flexible resources and providing these services to the transmission network. In the lower level, the prosumers act as price takers and change their supplies or demands based on the incentivizing prices sent by the aggregators.

The dominant scenario in the literature of game-theoretic algorithms in energy management systems is to consider a single aggregator and study the competition among different prosumers. In this work, on the other hand, we consider a multi-aggregator scheme in which the aggregators compete among each other to increase their profit in the market. A system operator aims to steer the aggregators towards the NE of the game. The challenge here is that the profit of an aggregator itself depends on the cost functions of all of its prosumers. The latter cost functions are unknown to the aggregator due to privacy reasons.

After carefully modeling the competition among aggrgeators and their coupling with the prosumers, we propose a fully distributed algorithm in discrete-time to solve the resulting optimization problem and steer the aggregators towards the NE of the game. Each aggregator only needs to exchange its estimate of the aggregate and an auxiliary variable with its neighbours.

Notation We use $\mathbf{1}$ to denote a vector of all ones. We use $\|A\|$ to denote the maximum singular value of A. For a differentiable scalar function f, we use f' and f^{-1} to denote its derivative and inverse function respectively.

Operator theoretic definitions We use $\mathrm{Id}(\cdot)$ to denote the identity operator. For a closed set $\Omega \in \mathbb{R}^n$, the mapping Proj_Ω denotes the projection onto Ω. The set-valued mapping N_Ω denotes the normal corn operator for the set $\Omega \in \mathbb{R}^n$. For a non-differentiable function f, ∂f denotes its subdifferential set-valued mapping, defined as $\partial f = \{v \in \mathbb{R}^n | f(z) \geq f(x) + v^\top (z - v)\}$ for all $z \in \mathrm{dom}(f)$. A set-valued mapping F is ℓ-Lipschitz continuous, with $\ell > 0$, if $\|u - v\| \leq \ell \|x - y\|$ for all $x, y \in \mathbb{R}^n$, $u \in \mathrm{F}(x)$, $v \in \mathrm{F}(y)$. The mapping F is μ-strongly monotone, with $\mu > 0$, if $(u - v)^\top)(x - y) \geq \mu \|x - y\|^2$ for all $x, y \in \mathbb{R}^n$, $u \in \mathrm{F}(x)$, $v \in \mathrm{F}(y)$. The mapping F is η-averaged, with $\eta \in (0, 1)$, if $\|u - v\|^2 \leq \ell \|x - y\|^2 - \frac{1-\eta}{\eta} \|(\mathrm{Id} - (F)(x)) - (\mathrm{Id} - (F)(y))\|^2$, for all $x, y \in \mathbb{R}^n$, $u \in \mathrm{F}(x)$, $v \in \mathrm{F}(y)$. The mapping F is β-cocoercive, with $\beta > 0$, if βF is $\frac{1}{2}$-averaged.

2. PROBLEM FORMULATION

We consider a scenario where there is an energy supply deficit in the wholesale electricity market. This can be due to a decrease in power generation from a wind or solar farm because of a change in weather conditions and power demand. The Transmission System Operator can procure energy flexibility services to complement this deficit from Distributed Energy Resources (DERs) connected at the distribution level. We design a fully decentralized energy management scheme for aggregating these DERs to freely participate in the wholesale market and introduce a group of aggregators to manage a population of prosumers, as shown in Fig.1. The case of an excess in energy supply can be treated analogously.

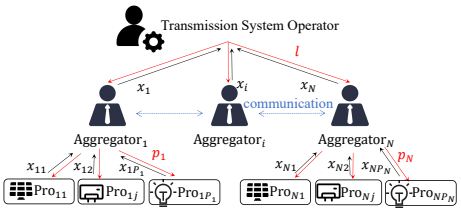


Fig. 1. The decentralized energy management scheme

2.1 Aggregator/Prosumer Model

The set of aggregators is denoted by $\mathcal{I} = \{1, 2, ..., N\}$ and aggregator i provides flexibility x_i by incentivizing the prosumers to consume less or generate more energy. Each aggregator aims to maximize its net revenue J_i, and we employ the following optimization problem,

$$\max_{0 \leq x_i \leq \bar{x}_i} J_i(x_i, p_i, s(\mathbf{x})) = -x_i p_i + l\big(s(\mathbf{x})\big) x_i \qquad (1)$$

where p_i and $l(\cdot)$ are the unit prices for buying and selling flexibility services respectively, $l : \mathbb{R} \to \mathbb{R}$, $s : \mathbb{R}^N \to \mathbb{R}$, and $\bar{x}_i$ is the maximum available flexibility due to the line capacity constraint. We assume that the selling price $l(\cdot)$ affinely depends on the aggregated flexibility, namely

$$l(s(\mathbf{x})) = l_0 - \lambda s(\mathbf{x}), \qquad (2)$$

where λ is a positive parameter corresponding to price elasticity and l_0 is a basic price for unit flexibility; $\mathbf{x} = \mathrm{col}(x_i)_{i \in \mathcal{I}}$, $s(\mathbf{x}) = \frac{1}{N} \mathbf{1}_N^\top \mathbf{x}$ and the coefficient $\frac{1}{N}$ is included for convenience. We note that the individual optimization problems are coupled with each other through the selling price l.

We denote the the prosumer ij's flexibility by x_{ij} for each $j \in \mathcal{P}_i = \{1, 2, ..., P_i\}$, where $\mathcal{P}_i$ is the set of prosumers assosiated with aggregator i. So the total flexibility $x_i := \sum_{j \in \mathcal{P}_i} x_{ij}$. The prosumer ij's goal is to maximize its revenue by altering its demand or supply based on the incentivizing price p_i. This results in the following optimization problem,

$$\max_{x_{ij} \geq 0} U_{ij}(x_{ij}) = x_{ij} p_i - g_{ij}(x_{ij}) \qquad (3)$$

where $U_{ij}(\cdot)$ denotes the utlitiy function of prosumer ij, and $g_{ij}(\cdot)$ accounts for the cost/inconvenience for providing the flexibility x_{ij}. We note that the first term in (3) is the payment received from aggregator i in return for the provided flexibility.

The model (1)-(3) for the aggregator i and its prosumers gives rise to a bilevel optimization problem. In the reminder of this section, we transfer this model into an equivalent convex problem by imposing a few assumptions.

Assumption 1. For each $j \in \mathcal{P}_i$ and $x_{ij} \geq 0$, the cost function g_{ij} is twice continuously differentiable and strictly convex with $g'_{ij}(0) \geq 0$. Moreover, we assume that g'_{ij} is convex, μ_{ij}-strongly monotone and ℓ_{ij}-Lipschitz continuous, for some constant $\mu_{ij}, \ell_{ij} > 0$.

The most common cost function, linear quadratic function $g_{ij} = \frac{1}{2} a_{ij} x_{ij}^2 + b_{ij} x_{ij}$, $a_{ij} > 0$, $b_{ij} > 0$, satisfies Assumption 1 with $\mu_{ij} = \ell_{ij} = a_{ij}$.

The prosumer maximization problem (3) admits the following well-known solution descending from the KKT condition.

Lemma 1. Let Assumption1 hold. Then, (3) admits a unique solution given by:

$$x_{ij} = \pi_{ij}(p_i) = \begin{cases} g_{ij}'^{-1}(p_i) & if\ p_i \geq g_{ij}'(0) \\ 0 & if\ p_i < g_{ij}'(0) \end{cases} \tag{4}$$

Then we can build the relationship between p_i and x_i by the following Lemma.

Lemma 2. Let Assumption 1 hold and let $\bar{\mathcal{P}}_i \subseteq \mathcal{P}_i$ denote the set of prosumers providing flexibility, i.e., $x_{ij} > 0$ for all $j \in \bar{\mathcal{P}}_i$. Then, there exits a continuous function u_i satisfying

$$p_i = u_i(x_i) \tag{5}$$

with the following properties: (i) $u_i \geq \max_{j\in M}\{g_{ij}'(0)\}$; (ii) u_i is continuous, strictly increasing and convex; (iii) μ_i-strongly monotone and ℓ_i-Lipschitz continuous, with $\mu_i = 1/\sum_{j\in\bar{\mathcal{P}}_i}\frac{1}{\mu_{ij}}$ and $\ell_i = 1/\sum_{j\in\bar{\mathcal{P}}_i}\frac{1}{\ell_{ij}}$.

To obtain more explicit expressions for optimal flexibility response x_i and price p_i, we take linear-quadratic function as an example for cost function and consider all prosumers entering the market. In this case, we have

$$u_i(x_i) = \beta_i + \alpha_i x_i \tag{6}$$

where $\alpha_i = 1/\sum_{j\in\mathcal{P}_i}\frac{1}{a_{ij}}$ and $\beta_i = \sum_{j\in\mathcal{P}_i}\frac{b_{ij}}{a_{ij}}/\sum_{j\in\mathcal{P}_i}\frac{1}{a_{ij}}$.

2.2 Game Theoretic Formulation

Following Lemma 2, we can substiute $p_i = u_i(x_i)$ in (5) to obtain

$$\min_{\underline{x}_i \leq x_i \leq \overline{x}_i} J_i(x_i, s(\mathbf{x})) = x_i u_i(x_i) - l\big(s(\mathbf{x})\big)x_i, \tag{7}$$

where $l\big(s(\mathbf{x})\big)$ is given by (2) and $\underline{x}_i := u_i^{-1}(\max_{j\in M}\{g_{ij}'(0)\})$.

We now write the noncooperative game among aggregators in a compact form as a triple

$$\mathcal{G} = \{\mathcal{I}, (\Omega_i)_{i\in\mathcal{I}}, (J_i(x_i, s(\mathbf{x})))_{i\in\mathcal{I}}\},$$

where $\mathcal{I}$ is the set of the aggregators participating in the game,

$$\Omega_i = \{x_i \in \mathbb{R} | \underline{x}_i \leq x_i \leq \overline{x}_i\}$$

is the set of possible strategies that aggregator i can take, and $J_i(x_i, s(\mathbf{x}))$ is the objective function. For this aggregative game, we obtain the subdifferential of the objective function with respect to x_i as

$$\begin{aligned} &f_i(x_i, s) = \partial_{x_i} J_i(x_i, s(\mathbf{x})) = \\ &u_i + (\partial_{x_i} u_i(x_i) + \lambda)x_i + N\lambda s - l_0. \end{aligned} \tag{8}$$

The following property plays a crucial role for our NE seeking algorithm design.

Lemma 3. Let Assumption 1 hold. Then, for all $i \in \mathcal{I}$, $x_i \in \Omega_i$ and $s \in \mathbb{R}$, the mapping $x_i \to f_i(x_i, s)$ is $\bar{\mu}_i$-strongly monotone with $\bar{\mu}_i := 2\mu_i + \lambda$ and $\bar{\ell}_i$-Lipschitz continuous with $\bar{\ell}_i := 2\ell_i + \lambda$.

The following lemma demonstrates the existence and uniqueness of the NE.

Lemma 4. Let Assumption 1 hold and $2\mu_i + \lambda > N\lambda$. Then, the aggregative game $\mathcal{G}$ has a unique NE $\mathbf{x}^*$, which satisfy

$$0 \in F(\mathbf{x}^*, s(\mathbf{x}^*)) + N_\Omega(\mathbf{x}^*) \tag{9}$$

where, $F(\mathbf{x}, s(\mathbf{x})) := \mathrm{col}(f_i(x_i, s(\mathbf{x}))_{i\in\mathcal{I}}$, $\Omega := \prod_{i\in\mathcal{I}}\Omega_i$.

We consider the aggaregators communicate only locally with their neighbours, over a communication graph G.

Assumption 2. The communication graph G is undirected and connected.

3. ALGORITHM

3.1 Algorithm Design

In this section we present our proposed algorithm. To offset the lack of full information, each aggregator i maintains a local estimate σ_i of the aggregate $s(\mathbf{x})$ and an additional auxiliary variable ψ_i to reach consensus of the local estimate, and exchanges them with its neighbours over G. The goal is that over time each aggregator will have the same aggregate estimate, equal to the average of the aggregators' actions, and its decision will correspond to NE. The proposed distributed algorithm is given below.

Algorithm 1 Distributed algorithm 1

Initialization: for all $i \in \mathcal{I}$, set $x_i(0) \in \Omega_i$, $\sigma_i(0) \in \mathbb{R}$, $\psi_i(0) \in \mathbb{R}$

Iterate until convergence: For all $i \in \mathcal{I}$, aggregator i exchanges $\sigma_i(k)$, $\psi_i(k)$ with its neighbours $\mathcal{N}_i$.

Local variable update:

$$\begin{aligned} x_i(k+1) &= \mathrm{Proj}_{\Omega_i}(x_i(k) - \tau_i\kappa_i f_i(x_i(k), \sigma_i(k))) \\ \psi_i(k+1) &= \psi_i(k) + v_i(\sum_{j\in\mathcal{N}_i}(\sigma_i(k) - \sigma_j(k))) \\ \sigma_i(k+1) &= \sigma_i(k) + \rho_i(-\sigma_i(k) + x_i(k) \\ &\quad - \sum_{j\in\mathcal{N}_i}(2\psi_i(k+1) - \psi_i(k))) \end{aligned}$$

where $\kappa_i > 0$ is a designed parameter, τ_i, v_i and ρ_i are positive step sizes.

To write the algorithm more compactly, let $\boldsymbol{\psi} = \mathrm{col}(\psi_i)_{i\in\mathcal{I}}$, $\boldsymbol{\sigma} = \mathrm{col}(\sigma_i)_{i\in\mathcal{I}}$, $K = \mathrm{diag}(\kappa_i)_{i\in\mathcal{I}}$, $\boldsymbol{\tau} = \mathrm{diag}(\tau_i)_{i\in\mathcal{I}}$, $\mathbf{v} = \mathrm{diag}(v_i)_{i\in\mathcal{I}}$, $\boldsymbol{\rho} = \mathrm{diag}(\rho_i)_{i\in\mathcal{I}}$, $F(\mathbf{x}, \boldsymbol{\sigma}) = \mathrm{col}(f_i(x_i, \sigma_i))_{i\in\mathcal{I}}$. Consequently, we can write the dynamics in Algorithm 1 as,

$$\begin{aligned} &\mathbf{x}(k+1) = \mathrm{Proj}_\Omega(\mathbf{x}(k) - \boldsymbol{\tau}KF(\mathbf{x}(k), \boldsymbol{\sigma}(k))) \\ &\boldsymbol{\psi}(k+1) = \boldsymbol{\psi}(k) + \mathbf{v}L\boldsymbol{\psi}(k) \\ &\boldsymbol{\sigma}(k+1) = \boldsymbol{\sigma}(k) + \boldsymbol{\rho}(\mathbf{x}(k) - \boldsymbol{\sigma}(k) - L(2\boldsymbol{\psi}(k+1) - \boldsymbol{\psi}(k))) \end{aligned} \tag{10}$$

where L is the Laplacian matrix of G.

3.2 Steady-State Analysis

Before we provide a convergence analysis, we show that the steady state of the dynamics in (10) yields the NE of the game. To this end, we define

$$\Gamma([\mathbf{x}; \boldsymbol{\psi}; \boldsymbol{\sigma}]) := \begin{bmatrix} KF(\mathbf{x}, \boldsymbol{\sigma}) + N_\Omega(\mathbf{x}) \\ -L\boldsymbol{\sigma} \\ \boldsymbol{\sigma} - \mathbf{x} + L\boldsymbol{\psi} \end{bmatrix} \tag{11}$$

In what follows, we show that (10) can be written as the following preconditioned forward-backward iteration:

$$\begin{aligned} \boldsymbol{\omega}(k+1) &= (\mathrm{Id} + \Phi^{-1}\mathcal{B})^{-1} \circ (\mathrm{Id} - \Phi^{-1}\mathcal{A})(\boldsymbol{\omega}(k)) \\ &= \mathcal{V}_\Phi \circ \mathcal{U}_\Phi \end{aligned} \tag{12}$$

where $\boldsymbol{\omega} = \text{col}(\mathbf{x}, \boldsymbol{\psi}, \boldsymbol{\sigma})$, $\mathcal{U}_\Phi = (\text{Id} - \Phi^{-1}\mathcal{A})$, $\mathcal{V}_\Phi = (\text{Id} + \Phi^{-1}\mathcal{B})^{-1}$, two operators $\mathcal{A}$ and $\mathcal{B}$ are split from Γ

$$\mathcal{A} := \begin{bmatrix} KF(\mathbf{x}, \boldsymbol{\sigma}) \\ 0 \\ -\mathbf{x} + \boldsymbol{\sigma} \end{bmatrix}, \mathcal{B} := \begin{bmatrix} N_\Omega(\mathbf{x}) \\ -L\boldsymbol{\sigma} \\ L\boldsymbol{\psi} \end{bmatrix} \tag{13}$$

and

$$\Phi = \begin{bmatrix} \boldsymbol{\tau}^{-1} & 0 & 0 \\ 0 & \mathbf{v}^{-1} & L \\ 0 & L & \boldsymbol{\rho}^{-1} \end{bmatrix}. \tag{14}$$

The main result of this subsection is provided below.

Proposition 1. Let Assumption 1 and 2 hold. Assume that $2\mu_i + \lambda > N\lambda$, τ_i, v_i and ρ_i are all positive, and let $\boldsymbol{\beta}$ and $\mathbf{v}$ be chosen such that

$$\max\{\rho_i\}_{i\in\mathcal{I}} < (\|L\|^2\mathbf{v})^{-1}. \tag{15}$$

Then, dynamics (10) is equivalent to the forward-backward iteration (12), hence the steady state $\overline{\boldsymbol{\omega}} = \text{col}(\overline{\mathbf{x}}, \overline{\boldsymbol{\psi}}, \overline{\boldsymbol{\sigma}})$ of (10) coincides with the fix point of iteration (12) and the zero of the mapping Γ in (11). Moreover, $\overline{\mathbf{x}}$ is the NE $\mathbf{x}^*$ of game $\mathcal{G}$.

3.3 Convergence Analysis

In this subsection, we show the convergence of Algorithm 1 to NE under suitable choices of step sizes.

We state several results which we will use to claim the convergence of Algorithm 1.

Lemma 5. Let Assumption1 and 2 hold, $2\mu_i + \lambda > N\lambda$ and

$$\kappa_i \in \left(\frac{(\sqrt{\bar{\mu}_i} - \sqrt{\bar{\mu}_i - N\lambda})^2}{\bar{\ell}_i^2}, \frac{(\sqrt{\bar{\mu}_i} + \sqrt{\bar{\mu}_i - N\lambda})^2}{\bar{\ell}_i^2} \right) \tag{16}$$

for $i \in \mathcal{I}$, the mapping $\widetilde{\mathcal{A}}$ is cocoercive,

$$\widetilde{\mathcal{A}} := \begin{bmatrix} KF(\mathbf{x}, \boldsymbol{\sigma}) \\ -\mathbf{x} + \boldsymbol{\sigma} \end{bmatrix} \tag{17}$$

Using the fact that $\tilde{\mathcal{A}}$ is cocoercive, we can now show properties for the operators $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{U}_\Phi$, $\mathcal{V}_\Phi$ in the Φ-induced norm.

Lemma 6. Let Assumption1 and 2 hold, $2\mu_i + \lambda > N\lambda$, κ_i satisfies (16), τ_i, v_i and ρ_i are all positive and satisfy (15). Then, $\Phi^{-1}\mathcal{A}$ is ξ-cocoercive, $\mathcal{U}_\Phi$ is $\frac{1}{2\xi}$-averaged, $\Phi^{-1}\mathcal{B}$ is maximally monotone, $\mathcal{V}_\Phi$ is $\frac{1}{2}$-averaged in the Φ-induced norm.

Next we show the operator $\mathcal{V}_\Phi \circ \mathcal{U}_\Phi$ is averaged if the step sizes are chosen small enough.

Lemma 7. The forward-backward iteration in (12), is θ-averaged, with $\theta = \frac{1}{2-1/(2\xi)} \in (0, 1)$, if

$$\tau_i < \frac{2\epsilon}{\tilde{\ell}^2}, \rho_i < \frac{2\epsilon}{\tilde{\ell}^2}, v_i < \frac{1}{\|L\|^2}\left(\frac{1}{\rho_i} - \frac{\tilde{\ell}^2}{2\epsilon}\right) \tag{18}$$

where $\tilde{\ell} = \max\{\max\{\bar{\ell}_i\}_{i\in\mathcal{I}}, N\lambda\} + 1$, $\epsilon = \min\{\epsilon_i\}_{i\in\mathcal{I}}$ and $\epsilon_i = -\frac{\sqrt{(N\lambda\kappa_i+1)^2+(\kappa_i\bar{\mu}_i-1)^2}}{2} + \frac{\kappa_i\bar{\mu}_i+1}{2}$.

Now we can show the convergence of Algorithm 1.

Theorem 1. Let Assumption1 and 2 hold, $2\mu_i+\lambda > N\lambda$, κ_i satisfies (16), $\boldsymbol{\tau}$, $\boldsymbol{\beta}$ and $\mathbf{v}$ are chosen as in Lemma 7. Then, the iteration (12) is θ-averaged, with $\theta \in (0, 1)$, thus the sequence defined by Algorithm 1 converges to the zero of the mapping Γ, $\overline{\boldsymbol{\omega}} = \text{col}(\overline{\mathbf{x}}, \overline{\boldsymbol{\psi}}, \overline{\boldsymbol{\sigma}})$.

Proof of Theorem 1(sketch): It follows by Proposition 1 that Algorithm 1 corresponds to the fix point of iteration (12) of the mapping $\mathcal{V}_\Phi \circ \mathcal{U}_\Phi$, which is θ-averaged, with $\theta \in (0, 1)$, by Lemma 7, if the step sizes satisfy (18). Then the convergence of the sequence generated by the iteration (12) to $\overline{\boldsymbol{\omega}} = \text{col}(\overline{\mathbf{x}}, \overline{\boldsymbol{\psi}}, \overline{\boldsymbol{\sigma}})$ follows by (Bauschke et al. (2011), Prop. 15.5).

REFERENCES

Bauschke, H.H., Combettes, P.L., et al. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer.

Belgioioso, G. and Grammatico, S. (2019). A distributed proximal-point algorithm for nash equilibrium seeking in generalized potential games with linearly coupled cost functions. In *2019 18th European Control Conference (ECC)*, 1–6. IEEE.

De Persis, C. and Grammatico, S. (2019). Continuous-time integral dynamics for a class of aggregative games with coupling constraints. *IEEE Transactions on Automatic Control*, 65(5), 2171–2176.

Franci, B. and Grammatico, S. (2020). A distributed forward–backward algorithm for stochastic generalized nash equilibrium seeking. *IEEE Transactions on Automatic Control*, 66(11), 5467–5473.

Gadjov, D. and Pavel, L. (2020). Single-timescale distributed gne seeking for aggregative games over networks via forward–backward operator splitting. *IEEE Transactions on Automatic Control*, 66(7), 3259–3266.

Gkatzikis, L., Koutsopoulos, I., and Salonidis, T. (2013). The role of aggregators in smart grid demand response markets. *IEEE Journal on selected areas in communications*, 31(7), 1247–1257.

Koshal, J., Nedić, A., and Shanbhag, U.V. (2016). Distributed algorithms for aggregative games on graphs. *Operations Research*, 64(3), 680–704.

Parvania, M., Fotuhi-Firuzabad, M., and Shahidehpour, M. (2013). Optimal demand response aggregation in wholesale electricity markets. *IEEE transactions on smart grid*, 4(4), 1957–1965.

Pavel, L. (2019). Distributed gne seeking under partial-decision information over networks via a doubly-augmented operator splitting approach. *IEEE Transactions on Automatic Control*, 65(4), 1584–1597.

Shakarami, M., De Persis, C., and Monshizadeh, N. (2019). Privacy and robustness guarantees in distributed dynamics for aggregative games. *arXiv preprint arXiv:1910.13928*.

Yi, P. and Pavel, L. (2018). Distributed generalized nash equilibria computation of monotone games via double-layer preconditioned proximal-point algorithms. *IEEE Transactions on Control of Network Systems*, 6(1), 299–311.

Zugno, M., Morales, J.M., Pinson, P., and Madsen, H. (2013). A bilevel model for electricity retailers' participation in a demand response market environment. *Energy Economics*, 36, 182–197.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A structure-preserving finite element method for MHD that preserves energy, cross-helicity, magnetic helicity, divB = 0 ⋆

François Gay-Balmaz * **Evan S. Gawlik** **

* *CNRS - Ecole Normale Supérieure, Paris, 24 Rue Lhomond (e-mail: francois.gay-balmaz@lmd.ens.fr).*
** *Department of Mathematics, University of Hawai'i at Manoa USA, (e-mail: egawlik@hawaii.edu)*

Abstract: We construct a structure-preserving finite element method and time-stepping scheme for inhomogeneous, incompressible magnetohydrodynamics (MHD). The method preserves energy, cross-helicity (when the fluid density is constant), magnetic helicity, mass, total squared density, pointwise incompressibility, and the constraint $\operatorname{div} B = 0$ to machine precision, both at the spatially and temporally discrete levels.

Keywords: Structure-preserving discretization, magnetohydrodynamics, finite element method, variational formulation, conservation laws.

1. INTRODUCTION

We construct a structure-preserving finite element method for solving the inhomogeneous, incompressible magneto-hydrodynamic (MHD) equations on a bounded domain $\Omega \subset \mathbb{R}^3$. These equations seek a velocity field u, magnetic field B, pressure p, and density ρ satisfying

$$\rho(\partial_t u + u \cdot \nabla u) - (\nabla \times B) \times B = -\nabla p, \text{ in } \Omega \times (0,T), \quad (1)$$
$$\partial_t B - \nabla \times (u \times B) = 0, \text{ in } \Omega \times (0,T), \quad (2)$$
$$\partial_t \rho + \operatorname{div}(\rho u) = 0, \text{ in } \Omega \times (0,T), \quad (3)$$
$$\operatorname{div} u = \operatorname{div} B = 0, \text{ in } \Omega \times (0,T), \quad (4)$$
$$u \cdot n = B \cdot n = 0, \text{ on } \partial\Omega \times (0,T), \quad (5)$$
$$u(0) = u_0,\ B(0) = B_0,\ \rho(0) = \rho_0, \text{ in } \Omega. \quad (6)$$

The method we construct exactly preserves energy $\frac{1}{2}\int_\Omega \rho u \cdot u + B \cdot B\, dx$, cross-helicity $\int_\Omega u \cdot B\, dx$ (when $\rho \equiv 1$), magnetic helicity $\int_\Omega A \cdot B\, dx$, mass $\int_\Omega \rho\, dx$, total squared density $\int_\Omega \rho^2\, dx$, and the constraints $\operatorname{div} u = \operatorname{div} B = 0$ at the spatially and temporally discrete level. Here, A denotes the magnetic potential; that is, A is any vector field satisfying $\nabla \times A = B$ and $A \times n|_{\partial\Omega} = 0$.

Our method, developed in Gawlik and Gay-Balmaz (2022), builds upon a growing body of literature on structure preservation in incompressible MHD simulations. Much of this literature focuses on the setting of constant density. In that setting, researchers have constructed energy-stable schemes that preserve $\operatorname{div} B = 0$ Hu et al. (2017); energy-stable schemes that preserve $\operatorname{div} u = \operatorname{div} B = 0$ Hiptmair et al. (2018); schemes that preserve energy, cross-helicity, and $\operatorname{div} u = \operatorname{div} B = 0$ Liu and Wang (2001); Gawlik et al. (2011); and schemes that preserve energy, cross-helicity, $\int A\, dx$, and $\operatorname{div} u = \operatorname{div} B = 0$ in two dimensions Kraus and Maj (2017). More recently, Hu, Lee, and Xu Hu et al. (2021) constructed a finite element method for homogeneous, incompressible MHD that preserves energy, cross-helicity, magnetic helicity, and $\operatorname{div} B = 0$.

⋆ Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lowercase letters, not all uppercase.

2. WEAK FORMULATION AND CONSERVED QUANTITIES

For every pair of smooth vector fields v and C satisfying $v \cdot n|_{\partial\Omega} = C \cdot n|_{\partial\Omega} = 0$ and every pair of smooth scalar fields σ and q, the solution (u, B, ρ, p) of (1-6) satisfies

$$\langle \partial_t(\rho u), v\rangle + a(\rho u, u, v) - a(B, B, v) + b(u \cdot u/2, \rho, v) = \langle \widetilde{p}, \operatorname{div} v\rangle, \quad (7)$$
$$\langle \partial_t B, C\rangle + a(C, B, u) = 0, \quad (8)$$
$$\langle \partial_t \rho, \sigma\rangle + b(\sigma, \rho, u) = 0, \quad (9)$$
$$\langle \operatorname{div} u, q\rangle = 0, \quad (10)$$

where $\widetilde{p} = p + \rho u \cdot u$ and

$$a(w, u, v) = \langle w, \nabla \times (u \times v)\rangle,$$
$$b(f, g, w) = -\langle w \cdot \nabla f, g\rangle.$$

The structure of equations (7-10) is made even more transparent if one introduces the Lagrangian $\ell(u, B, \rho) = \frac{1}{2}\langle \rho u, u\rangle - \frac{1}{2}\langle B, B\rangle$ of inhomogeneous, incompressible MHD. In terms of $\frac{\delta \ell}{\delta u} = \rho u$, $\frac{\delta \ell}{\delta B} = -B$, and $\frac{\delta \ell}{\delta \rho} = \frac{1}{2} u \cdot u$, equations (7-10) take the form

$$\Big\langle \partial_t \frac{\delta \ell}{\delta u}, v\Big\rangle + a\Big(\frac{\delta \ell}{\delta u}, u, v\Big) + a\Big(\frac{\delta \ell}{\delta B}, B, v\Big) + b\Big(\frac{\delta \ell}{\delta \rho}, \rho, v\Big) = \langle \widetilde{p}, \operatorname{div} v\rangle, \quad (11)$$
$$\langle \partial_t B, C\rangle + a(C, B, u) = 0, \quad (12)$$
$$\langle \partial_t \rho, \sigma\rangle + b(\sigma, \rho, u) = 0, \quad (13)$$
$$\langle \operatorname{div} u, q\rangle = 0. \quad (14)$$

It is this variational structure that inspired the numerical method we propose, see Gawlik and Gay-Balmaz (2021a,b, 2022).

The formulation (7-10) allows one to easily deduce its invariants of motion from basic properties of the trilinear forms a and b. Namely, a and b satisfy:

$$a(w,u,v) = -a(w,v,u), \tag{15}$$
$$a(w,u,v) = 0 \text{ if } u\cdot n|_{\partial\Omega} = v\cdot n|_{\partial\Omega} = 0, \nabla\times w = u, \tag{16}$$
$$b(f,g,w) = -b(g,f,w) \text{ if } \operatorname{div} w = 0,\ w\cdot n|_{\partial\Omega} = 0. \tag{17}$$

These properties give rise to the following conservation laws. Taking $\sigma = 1$ in (9) gives:

$$\frac{d}{dt}\int_\Omega \rho\,dx = \langle \partial_t\rho, 1\rangle = -b(1,\rho,u) = 0.$$

Taking $\sigma = \rho$ in (9) and using (17) gives:

$$\frac{d}{dt}\frac{1}{2}\int_\Omega \rho^2\,dx = \langle \partial_t\rho,\rho\rangle = -b(\rho,\rho,u) = 0.$$

Taking $v = u$ in (7) and $C = B$ in (8) gives:

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_\Omega \rho u\cdot u + B\cdot B\,dx\\ &\quad= \langle\partial_t(\rho u), u\rangle - \langle\partial_t\rho, u\cdot u/2\rangle + \langle\partial_t B, B\rangle\\ &\quad= \langle\widetilde{p}, \operatorname{div} u\rangle - a(\rho u,u,u) + a(B,B,u)\\ &\qquad - b(u\cdot u/2,\rho,u) - \langle\partial_t\rho, u\cdot u/2\rangle - a(B,B,u) = 0,\end{aligned}$$

where we used $\operatorname{div} u = 0$, (15), and (9). If A satisfies $\nabla\times A = B$ and $A\times n|_{\partial\Omega} = 0$, then:

$$\begin{aligned}&\frac{d}{dt}\int_\Omega A\cdot B\,dx = \langle\partial_t A, B\rangle + \langle A,\partial_t B\rangle\\ &\quad= \langle\partial_t A, \nabla\times A\rangle + \langle A,\partial_t B\rangle = \langle\nabla\times(\partial_t A), A\rangle + \langle A,\partial_t B\rangle\\ &\quad= \langle\partial_t B, A\rangle + \langle A,\partial_t B\rangle = -2a(A,B,u) = 0,\end{aligned}$$

from (8) and (16). Finally, if $\rho\equiv 1$, then taking $v = B$ in (7) and $C = u$ in (8) gives:

$$\begin{aligned}\frac{d}{dt}\int u\cdot B\,dx &= \langle\partial_t u, B\rangle + \langle\partial_t B, u\rangle\\ &= \langle\widetilde{p}, \operatorname{div} B\rangle - a(u,u,B) + a(B,B,B)\\ &\quad - b(u\cdot u/2, 1, B) - a(u,B,u) = 0,\end{aligned}$$

where we used $\operatorname{div} B = 0$, $b(u\cdot u,1,B) = -b(1,u\cdot u,B) = 0$, and (15).

3. SPATIAL DISCRETIZATION

To construct a spatial discretization of (7-10) that preserves all the invariants discussed in Section 2, we will design discretizations of the trilinear forms a and b that satisfy analogues of (15), (16), and (17). By a careful choice of finite element spaces, the method we construct will also preserve the constraints $\operatorname{div} u = 0$ and $\operatorname{div} B = 0$ pointwise.

We make use of the following function spaces:

$$\begin{aligned}H_0^1(\Omega) &= \{f\in L^2(\Omega)\mid \nabla f\in L^2(\Omega)^3,\ f = 0 \text{ on } \partial\Omega\},\\ H_0(\operatorname{curl},\Omega) &= \{u\in L^2(\Omega)^3\mid \operatorname{curl} u\in L^2(\Omega)^3,\ u\times n = 0 \text{ on } \partial\Omega\}\\ H_0(\operatorname{div},\Omega) &= \{u\in L^2(\Omega)^3\mid \operatorname{div} u\in L^2(\Omega),\ u\cdot n = 0 \text{ on } \partial\Omega\},\\ \mathring{H}(\operatorname{div},\Omega) &= \{u\in H_0(\operatorname{div},\Omega)\mid \operatorname{div} u = 0\},\\ L^2_{\int=0}(\Omega) &= \{f\in L^2(\Omega)\mid \textstyle\int_\Omega f\,dx = 0\}.\end{aligned}$$

Let $\mathcal{T}_h$ be a triangulation of Ω, and let $\mathcal{E}_h$ denote the set of interior 2-dimensional faces in $\mathcal{T}_h$. For each integer $s\geq 0$ and each simplex $K\in\mathcal{T}_h$, we denote by $P_s(K)$ the space of polynomials of degree at most s on K. On a face $e = K_1\cap K_2\in\mathcal{E}_h$, we denote the jump and average of a piecewise smooth scalar function f by

$$[\![f]\!] = f_1 n_1 + f_2 n_2,\quad \{f\} = \frac{f_1+f_2}{2},$$

where $f_i = f|_{K_i}$, n_1 is the normal vector to e pointing from K_1 to K_2, and similarly for n_2.

Our numerical method will make use of four approximation spaces: a space $U_h^{\operatorname{div}}\subset H_0(\operatorname{div},\Omega)$ for the velocity u and magnetic field B, a space $F_h\subset L^2(\Omega)$ for the density ρ, a space $Q_h\subset L^2_{\int=0}(\Omega)$ for the pressure $\widetilde{p}$, and an auxiliary space $U_h^{\operatorname{curl}}\subset H_0(\operatorname{curl},\Omega)$. For the velocity and magnetic field, we use the Raviart-Thomas space

$$\begin{aligned}RT_s(\mathcal{T}_h) = \{&u\in H_0(\operatorname{div},\Omega)\mid\\ &u|_K\in P_s(K)^3 + xP_s(K), \forall K\in\mathcal{T}_h\},\end{aligned}$$

where $s\geq 0$ is an integer. For the pressure, we use the zero-mean subspace of the discontinuous Galerkin space

$$DG_s(\mathcal{T}_h) = \{f\in L^2(\Omega)\mid f|_K\in P_s(K),\ \forall K\in\mathcal{T}_h\}.$$

For the density, we use $DG_m(\mathcal{T}_h)$, where $m\geq 0$ is an integer (not necessarily equal to s). For the auxiliary space $U_h^{\operatorname{curl}}$, we use the space of Nedelec elements of the first kind,

$$\begin{aligned}NED_s(\mathcal{T}_h) = \{&u\in H_0(\operatorname{curl},\Omega)\mid\\ &u|_K\in P_s(K)^3 + x\times P_s(K)^3,\ \forall K\in\mathcal{T}_h\}.\end{aligned}$$

In summary,

$$U_h^{\operatorname{div}} = RT_s(\mathcal{T}_h),\qquad Q_h = DG_s(\mathcal{T}_h)\cap L^2_{\int=0}(\Omega),$$
$$U_h^{\operatorname{curl}} = NED_s(\mathcal{T}_h),\qquad F_h = DG_m(\mathcal{T}_h).$$

We define trilinear forms $a_h : L^2(\Omega)^3\times L^4(\Omega)^3\times L^4(\Omega)^3\to\mathbb{R}$ and $b_h : L^2(\Omega)\times L^2(\Omega)\times U_h^{\operatorname{div}}\to\mathbb{R}$ by

$$a_h(w,u,v) = \int_\Omega w\cdot\nabla\times\pi_h^{\operatorname{curl}}(\pi_h^{\operatorname{curl}}u\times\pi_h^{\operatorname{curl}}v)\,dx, \tag{18}$$

$$b_h(f,g,u) = -\sum_{K\in\mathcal{T}_h}\int_K (u\cdot\nabla\pi_h f)\pi_h g\,dx \tag{19}$$

$$+\sum_{e\in\mathcal{E}_h}\int_e u\cdot[\![\pi_h f]\!]\{\pi_h g\}\,ds, \tag{20}$$

where $\pi_h^{\operatorname{curl}} : L^2(\Omega)^3\to U_h^{\operatorname{curl}}$ and $\pi_h : L^2(\Omega)\to F_h$ denote the L^2-orthogonal projectors onto $U_h^{\operatorname{curl}}$ and F_h, respectively. Note that b_h (restricted to $F_h\times F_h\times U_h^{\operatorname{div}}$) is a standard discontinous Galerkin discretization of the scalar advection operator Brezzi et al. (2004).

These trilinear forms possess two important properties that mimic (15-17). The trilinear form a_h is alternating in its last two arguments:

$$a_h(w,u,v) = -a_h(w,v,u), \tag{21}$$

for all $(w,u,v)\in L^2(\Omega)^3\times L^4(\Omega)^3\times L^4(\Omega)^3$. Second, using integration by parts, one checks that b_h is alternating in its first two arguments if its last argument is divergence-free:

$$b_h(f,g,u) = -b_h(g,f,u), \tag{22}$$

for all $(f,g,u)\in L^2(\Omega)\times L^2(\Omega)\times(U_h^{\operatorname{div}}\cap\mathring{H}(\operatorname{div},\Omega))$. The additional property of a_h is shown as follows.

Lemma 3.1. The trilinear form (18) satisfies

$$a_h(w,u,v) = 0 \text{ if } \nabla\times w = u. \tag{23}$$

Proof. If $\nabla\times w = u$, then we can integrate (18) by parts and use the fact that $n\times\pi_h^{\operatorname{curl}}(\pi_h^{\operatorname{curl}}u\times\pi_h^{\operatorname{curl}}v)\big|_{\partial\Omega} = 0$ to obtain

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$$\begin{aligned} a_h(w,u,v) &= \langle w, \nabla \times \pi_h^{\mathrm{curl}}(\pi_h^{\mathrm{curl}} u \times \pi_h^{\mathrm{curl}} v)\rangle \\ &= \langle \nabla \times w, \pi_h^{\mathrm{curl}}(\pi_h^{\mathrm{curl}} u \times \pi_h^{\mathrm{curl}} v)\rangle \\ &= \langle u, \pi_h^{\mathrm{curl}}(\pi_h^{\mathrm{curl}} u \times \pi_h^{\mathrm{curl}} v)\rangle \\ &= \langle \pi_h^{\mathrm{curl}} u, \pi_h^{\mathrm{curl}} u \times \pi_h^{\mathrm{curl}} v\rangle = 0. \quad \blacksquare \end{aligned}$$

We define our semidiscrete numerical method as follows. We seek $u, B \in U_h^{\mathrm{div}}$, $\rho \in F_h$, and $p \in Q_h$ such that

$$\begin{aligned} \langle \partial_t(\rho u), v\rangle + a_h(\rho u, u, v) - a_h(B,B,v) \\ + b_h(u \cdot u/2, \rho, v) = \langle p, \operatorname{div} v\rangle, \quad (24) \\ \langle \partial_t B, C\rangle + a_h(C, B, u) = 0, \quad (25) \\ \langle \partial_t \rho, \sigma\rangle + b_h(\sigma, \rho, u) = 0, \quad (26) \\ \langle \operatorname{div} u, q\rangle = 0, \quad (27) \end{aligned}$$

for all $v \in U_h^{\mathrm{div}}, C \in U_h^{\mathrm{div}}, \sigma \in F_h, q \in Q_h$.

Proposition 3.2. (Gawlik and Gay-Balmaz (2022)). The numerical method (24-27) exactly preserves $\int_\Omega \rho\, dx$, $\int_\Omega \rho^2\, dx$, $\int_\Omega \rho u \cdot u + B \cdot B\, dx$, $\int_\Omega A \cdot B\, dx$, and (if $\rho \equiv 1$) $\int_\Omega u \cdot B\, dx$. Furthermore, $\operatorname{div} u(t) \equiv 0$ and $\operatorname{div} B(t) \equiv 0$ for every t.

Proof. We only prove the conservation of magnetic helicity, see Gawlik and Gay-Balmaz (2022) for the other ones. From (23), if A is any vector field satisfying $\nabla \times A = B$ and $A \times n|_{\partial\Omega} = 0$, then

$$\begin{aligned} \frac{d}{dt}\langle A, B\rangle &= \langle \partial_t A, B\rangle + \langle A, \partial_t B\rangle \\ &= \langle \partial_t A, \nabla \times A\rangle + \langle A, \partial_t B\rangle \\ &= \langle \nabla \times \partial_t A, A\rangle + \langle A, \partial_t B\rangle \\ &= 2\langle \partial_t B, A\rangle = 2\langle \partial_t B, \pi_h^{\mathrm{div}} A\rangle \\ &= -2a_h(\pi_h^{\mathrm{div}} A, B, u) = -2a_h(A,B,u) = 0. \quad (28) \end{aligned}$$

Above, we used (25) with $C = \pi_h^{\mathrm{div}} A$, and we used the fact that $a_h(\pi_h^{\mathrm{div}} A, B, u) = \langle \pi_h^{\mathrm{div}} A, \nabla \times \pi_h^{\mathrm{curl}}(\pi_h^{\mathrm{curl}} B \times \pi_h^{\mathrm{curl}} u)\rangle = a_h(A,B,u)$ since $\nabla \times U_h^{\mathrm{curl}} \subseteq U_h^{\mathrm{div}}$. $\blacksquare$

4. UPWINDING

To incorporate upwinding into the density advection equation (26), one can replace (26) by

$$\langle \partial_t \rho, \sigma\rangle + \widetilde{b}_h(u; \sigma, \rho, u) = 0, \forall \sigma \in F_h, \quad (29)$$

where we introduced the u-dependent trilinear form

$$\begin{aligned} &\widetilde{b}_h(u; f, g, v) \\ &= b_h(f,g,v) + \sum_{e \in \mathcal{E}_h} \int_e \beta_e(u) \left(\frac{v \cdot n}{u \cdot n}\right) [\![\pi_h f]\!] \cdot [\![\pi_h g]\!]\, ds. \quad (30) \end{aligned}$$

Here $\{\beta_e\}_{e \in \mathcal{E}_h}$ are nonnegative parameters which may depend on u. A standard choice for β_e is

$$\beta_e(u) = c|u \cdot n|,$$

Brezzi et al. (2004), where $c \in [0, \frac{1}{2}]$, although we have found that the smooth approximation

$$\beta_e(u) = \frac{2c}{\pi}(u \cdot n) \arctan\left(\frac{u \cdot n}{\varepsilon}\right) \quad (31)$$

with $\varepsilon > 0$ small (e.g. $\varepsilon = 0.01$) tends to give better numerical performance in our experiments. Full upwinding corresponds to the choice $c = \frac{1}{2}$ Brezzi et al. (2004). When $c > 0$, this modification of the density advection equation interferes with conservation of $\int_\Omega \rho^2\, dx$ and $\int_\Omega \rho u \cdot u + B \cdot B\, dx$, but not $\int_\Omega \rho\, dx$ since $\sum_{e \in \mathcal{E}_h} \int_e \beta_e(u) [\![1]\!] \cdot [\![\rho]\!]\, ds = 0$. However, there is a simple way to restore energy conservation. As suggested in Gawlik and Gay-Balmaz (2020), one replaces the momentum equation (24) by

$$\langle \partial_t(\rho u), v\rangle + \langle \alpha, v\rangle + \widetilde{b}_h(u; \theta, \rho, v) = \langle p, \operatorname{div} v\rangle, \quad (32)$$

for all $v \in U_h^{\mathrm{div}}$.

Proposition 4.1. (Gawlik and Gay-Balmaz (2022)). With the exception of $\int_\Omega \rho^2\, dx$, all of the invariants listed in Proposition 3.2 are preserved by (24-27) if one replaces (24) and (26) by (29) and (32), respectively.

5. TEMPORAL DISCRETIZATION

We now describe a temporal discretization of (the upwinded version of) (24-27) that exactly preserves all of the original invariants of (the upwinded version of) (24-27).

We use a time step $\Delta t > 0$, and we write u_k to denote the value of the discrete solution u at time $t_k = k\Delta t$. We denote $u_{k+1/2} = (u_k + u_{k+1})/2$, with similar notation for p, B, ρ, and

$$(\rho u)_{k+1/2} = \frac{\rho_k u_k + \rho_{k+1} u_{k+1}}{2}.$$

We consider the time discretization

$$\begin{aligned} \langle D_{\Delta t}(\rho u), v\rangle + a_h(\rho u, u, v) - a_h(B,B,v) \\ + \widetilde{b}_h(u; u_k \cdot u_{k+1}/2, \rho, v) - \langle p_{k+1}, \operatorname{div} v\rangle = 0, \quad (33) \\ \langle D_{\Delta t} B, C\rangle + a_h(C, B, u) = 0, \quad (34) \\ \langle D_{\Delta t}\rho, \sigma\rangle + \widetilde{b}_h(u; \sigma, \rho, u) = 0, \quad (35) \\ \langle \operatorname{div} u_{k+1}, q\rangle = 0, \quad (36) \end{aligned}$$

for all $v \in U_h^{\mathrm{div}}$, $C \in U_h^{\mathrm{div}}$, $\sigma \in F_h$, $q \in Q_h$, where we abbreviate $u_{k+1/2}$, $B_{k+1/2}$, $\rho_{k+1/2}$, and $(\rho u)_{k+1/2}$ as u, B, ρ, and ρu, respectively, and use the notation $D_{\Delta t}(\rho u) = \frac{\rho_{k+1} u_{k+1} - \rho_k u_k}{\Delta t}$, $D_{\Delta t} B = \frac{B_{k+1} - B_k}{\Delta t}$, etc.

Proposition 5.1. (Gawlik and Gay-Balmaz (2022)). If B_0 satisfies $\operatorname{div} B_0 \equiv 0$, then the solution of (33-36) satisfies

$$\begin{aligned} \int_\Omega \rho_{k+1}\, dx &= \int_\Omega \rho_k\, dx, \quad (37) \\ \int_\Omega \rho_{k+1}^2\, dx &\le \int_\Omega \rho_k^2\, dx, \ (= \text{if } \beta_e = 0, \forall e \in \mathcal{E}_h), \quad (38) \\ \int_\Omega \rho_{k+1} u_{k+1} \cdot u_{k+1} + B_{k+1} \cdot B_{k+1}\, dx & \quad (39) \\ = \int_\Omega \rho_k u_k \cdot u_k + B_k \cdot B_k\, dx, & \quad (40) \\ \int_\Omega u_{k+1} \cdot B_{k+1}\, dx &= \int_\Omega u_k \cdot B_k\, dx, \ \text{if } \rho_0 \equiv 1, \quad (41) \\ \int_\Omega A_{k+1} \cdot B_{k+1}\, dx &= \int_\Omega A_k \cdot B_k\, dx, \quad (42) \end{aligned}$$

as well as $\operatorname{div} u_k \equiv 0$ and $\operatorname{div} B_k \equiv 0$, for every k. Here, A_k denotes any vector field satisfying $\nabla \times A_k = B_k$ and $A_k \times n|_{\partial\Omega} = 0$.

6. NUMERICAL EXAMPLE

We simulated a magnetohydrodynamic rotor on $\Omega = [0,1] \times [0,1]$ with initial conditions

$$\begin{aligned} u(x,y,0) &= g(r)\,(5 - 10y, 10x - 5)\,, \quad (43) \\ B(x,y,0) &= \left(\frac{5}{4\sqrt{\pi}}, 0\right), \quad (44) \\ \rho(x,y,0) &= 1 + 9g(r), \quad (45) \end{aligned}$$

where $r = \sqrt{(x-0.5)^2 + (y-0.5)^2}$ and

$$g(r) = \begin{cases} 1, & \text{if } r \leq 0.1, \\ (23 - 200r)/3, & \text{if } 0.1 < r < 0.115, \\ 0, & \text{if } r \geq 0.115. \end{cases}$$

This setup, which leads to the development of torsional Alfven waves, was considered in (Hiptmair and Pagliantini, 2018, Section 4.3.4) to test a numerical scheme for compressible MHD; here we test our scheme for inhomogeneous, incompressible MHD.

We imposed periodic boundary conditions in the x-direction and $u \cdot n = B \cdot n = 0$ along $y = 0$ and $y = 1$. We used (33-36) with a time step $\Delta t = 0.005$, polynomial degree $s = 0$, and a uniform triangulation of Ω with maximum element diameter $h = 2^{-6}\sqrt{2}$. We used upwinding for both the density and momentum.

Plots of the computed solution at time $t = 0.5$ are shown in Figure 1. Torsional Alfven waves are seen propagating away from the center of the domain in the horizontal direction. One can compare Figure 1 loosely with Figure 4.12 in Hiptmair and Pagliantini (2018), bearing in mind that the fluid considered there is compressible, and the solutions computed there are plotted at different times.

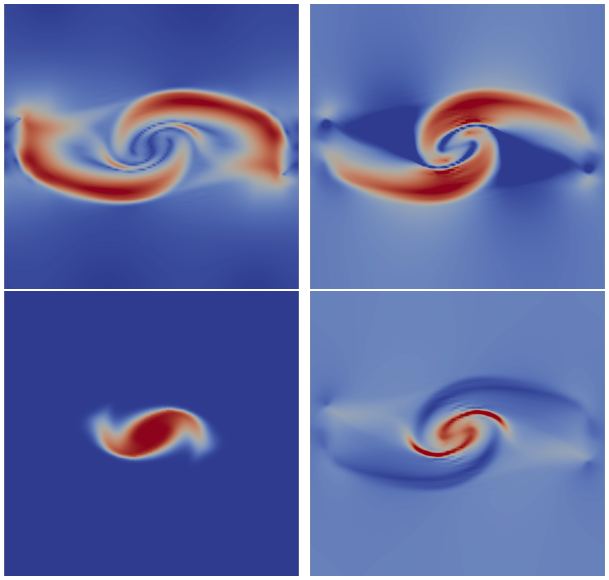


Fig. 1. Contours of $|u|$ (top left), $|B|$ (top right), ρ (down left), and p (down right) at time $t = 0.5$ in the magnetohydrodynamic rotor simulation.

REFERENCES

Brezzi, F., Marini, L.D., and Süli, E. (2004). Discontinuous Galerkin methods for first-order hyperbolic problems. *Mathematical Models and Methods in Applied Sciences*, 14(12), 1893–1903.

Gawlik, E.S. and Gay-Balmaz, F. (2021a). A variational finite element discretization of compressible flow. *Foundations of Computational Mathematics*, 21, 961–1001.

Gawlik, E.S. and Gay-Balmaz, F. (2020). A conservative finite element method for the incompressible euler equations with variable density. *Journal of Computational Physics*, 412, 109439.

Gawlik, E.S. and Gay-Balmaz, F. (2021b). A structure-preserving finite element method for compressible ideal and resistive MHD. *Journal of Plasma Physics*, 87(5).

Gawlik, E.S. and Gay-Balmaz, F. (2022). A finite element method for MHD that preserves energy, cross-helicity, magnetic helicity, incompressibility, and div B= 0. *Journal of Computational Physics*, 450(110847).

Gawlik, E.S., Mullen, P., Pavlov, D., Marsden, J.E., and Desbrun, M. (2011). Geometric, variational discretization of continuum theories. *Physica D: Nonlinear Phenomena*, 240(21), 1724–1760.

Hiptmair, R. and Pagliantini, C. (2018). Splitting-based structure preserving discretizations for magnetohydrodynamics. *SMAI Journal of Computational Mathematics*, 4, 225–257.

Hiptmair, R., Li, L., Mao, S., and Zheng, W. (2018). A fully divergence-free finite element method for magnetohydrodynamic equations. *Mathematical Models and Methods in Applied Sciences*, 28(04), 659–695.

Hu, K., Lee, Y.J., and Xu, J. (2021). Helicity-conservative finite element discretization for incompressible mhd systems. *Journal of Computational Physics*, 436, 110284.

Hu, K., Ma, Y., and Xu, J. (2017). Stable finite element methods preserving $\nabla \cdot B = 0$ exactly for mhd models. *Numerische Mathematik*, 135(2), 371–396.

Kraus, M. and Maj, O. (2017). Variational integrators for ideal magnetohydrodynamics. *arXiv preprint arXiv:1707.03227*.

Liu, J.G. and Wang, W.C. (2001). An energy-preserving MAC–Yee scheme for the incompressible MHD equation. *Journal of Computational Physics*, 174(1), 12–37.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Curse-of-Dimensionality-Free Computation of Control Lyapunov Functions Using Neural Networks ⋆

Lars Grüne * **Mario Sperl** *

* *Chair of Applied Mathematics, Mathematical Institute, University of Bayreuth, 95440 Bayreuth, Germany*

Keywords: deep neural network, curse of dimensionality, compositional function, control lyapunov function, nonlinear control system

1. INTRODUCTION

Control Lyapunov functions are a commonly used tool for studying stability and for constructing stabilizing feedback laws in systems and control theory. Since in most cases there is no analytic method to compute control Lyapunov functions, we are interested in the numerical computation of such functions. However, common numerical methods often suffer from the curse of dimensionality, i.e., an exponential growth of the computational effort in the state dimension. These approaches are thus limited to low-dimensional systems.

It is known that functions with certain beneficial structures can be approximated by deep neural networks without suffering from the curse of dimensionality. In this talk, we discuss the use of deep neural networks for an efficient approximation of control Lyapunov functions. To this end, we extend the approach for computing Lyapunov functions via deep neural networks that has been presented in Grüne (2021).

Our study is inspired by results concerning the approximation of solutions of partial differential equations and optimal value functions, see, e.g., Han et al. (2018); Darbon et al. (2020). These references rely on ideas from reinforcement learning. Alternatively, in Kang et al. (2021) a supervised learning based method is discussed, where an external algorithm provides values of a Lyapunov function that are then used for the training process. We note that while in our numerical examples we also use ideas from reinforcement learning, the complexity analysis presented in this talk is independent of the concrete learning method.

There exist several algorithmically orientated papers that use neural networks for the computation of control Lyapunov functions, see, e.g., Long and Bayoumi (1993); Khansari-Zadeh and Billard (2014); Richards et al. (2018). Moreover, the authors in Gaby et al. (2021) propose a training algorithm for a neural network that is based on a preliminary chosen Lyapunov function candidate. However, these papers do not provide an analytical investigation concerning the curse of dimensionality.

⋆ This research has been supported by the German Research Foundation (DFG) under project GR 1569/23-1 within the priority program 2298 "Theoretical Foundations of Deep Learning".

2. PROBLEM FORMULATION

We consider a nonlinear control system of the form

$$\dot{x}(t) = f(x(t), u(t)), \tag{1}$$

where $u \in L_\infty(\mathbb{R}, U)$ with $U \subset \mathbb{R}^m$ and $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous and Lipschitz continuous in x. We assume that the system (1) has an equilibrium at the origin, i.e., $f(0,0) = 0$.

Definition 1. Let $0 \in D \subset \mathbb{R}^n$ be open. A $\mathcal{C}^1$-function $V : D \to \mathbb{R}$ is called (smooth) control Lyapunov function if there exist $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{K}_\infty$ [1] such that

$$\alpha_1(\|x\|) \le V(x) \le \alpha_2(\|x\|), \tag{2}$$

$$\inf_{u \in U} DV(x) f(x,u) \le -\alpha_3(\|x\|) < 0 \tag{3}$$

for all $x \in D \setminus \{0\}$.

In the following, we assume that such a smooth control Lyapunov function exists for the system (1). Note that this is a sufficient condition for asymptotic null-controllability of (1) (cf. Sontag (1983)) and is in fact equivalent to the existence of a (possibly discontinuous) stabilizing feedback that is robust with respect to perturbations in the state (cf. Ledyaev and Sontag (1999)).

It is our goal to construct a neural network architecture that computes a control Lyapunov function avoiding the curse of dimensionality.

3. DEEP NEURAL NETWORKS

We want to briefly recall the concept of deep neural networks and compare the well-known universal approximation theorem to an approximation result for compositional functions.

Neural networks take an input vector and process it through a certain number of layers in order to produce an output. For our purpose of representing a control Lyapunov function, we use the input vector $x \in \mathbb{R}^n$ and a one-dimensional output $W(x;\theta) \in \mathbb{R}$. This means that the input layer possesses $N_0 = n$ neurons and the output layer consists of only one neuron. The numbers N_l of neurons in the remaining layers (called hidden layers) may vary.

[1] We define $\mathcal{K}_\infty$ as the space of all continuous and strictly increasing functions $\alpha\colon [0,\infty) \to [0,\infty)$ with $\alpha(0) = 0$ and $\lim_{\tau\to\infty} \alpha(\tau) = \infty$.

The vector $\theta \in \mathbb{R}^p$ represents parameters that determine the output of the neural network and have to be learned during the training process. Denote with $y_k^l \in \mathbb{R}$ the value of the neuron at position k in layer l. It is determined by the values of the neurons at the previous layer via

$$y_k^l = \sigma^l \left(\sum_{i=1}^{N_{l-1}} w_{k,i}^l \, y_i^{l-1} + b_k^l \right),$$

where $\sigma^l \colon \mathbb{R} \to \mathbb{R}$ is the so-called activation function of the l-th layer and $w_{k,i}^l$, $b_k^l \in \mathbb{R}$ are parameters that are comprised in θ. The activation function of the output layer is usually chosen to be an affine function. Figure 1 shows a neural network with one hidden layer, where the activation of the neuron y_1^1 is highlighted. Note that in our setting we have $y_i^0 = x_i$ for $i = 1, \dots, n$ and $y_1^2 = W(x; \theta)$.

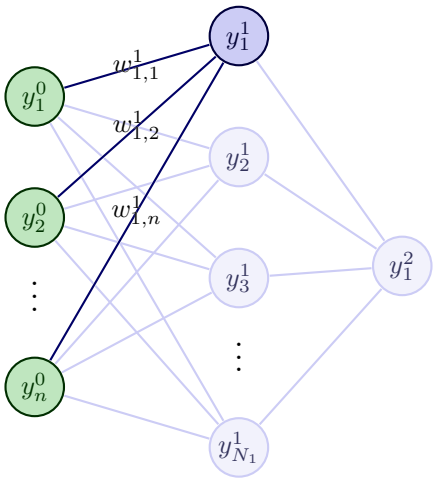


Fig. 1. Neural network with one hidden layer

Based on Mhaskar (1996), the following universal approximation theorem for single-layer neural networks has been proven in Poggio et al. (2017).

Theorem 2. Let $C > 0$ and $K_n \subset [-C, C]^n \subset \mathbb{R}^n$ be compact. Define

$$W_r^n := \left\{ g \in \mathcal{C}^r(K_n, \mathbb{R}) \,\middle|\, \sum_{1 \le |\alpha| \le r} \|D_\alpha g\|_{\infty, K} \le 1 \right\},$$

where for a continuous $g \colon K_n \to \mathbb{R}$ we set

$$\|g\|_{\infty, K} := \max_{x \in K_n} |g(x)|.$$

Moreover, let $\sigma^1 \colon \mathbb{R} \to \mathbb{R}$ be a $\mathcal{C}^\infty$-function that is not a polynomial. Then for any $\epsilon > 0$ there exists a number $N_\epsilon \in \mathbb{N}$ such that a neural network with one hidden layer that has at least N_ϵ neurons satisfies for all $g \in W_r^n$

$$\inf_\theta \|W(x; \theta) - g(x)\|_{\infty, K_n} \le \epsilon.$$

For the number of neurons it holds that

$$N_\epsilon = \mathcal{O}(\epsilon^{-\frac{n}{r}})$$

and this is best possible.

Theorem 2 states that neural networks with one hidden layer are capable of approximating $\mathcal{C}^1$-functions on compact sets. However, the required size of neurons still grows exponentially in the state dimension. For the class of compositional functions, the curse of dimensionality can be avoided.

Definition 3. A function $g \colon \mathbb{R}^n \to \mathbb{R}$ is called compositional of degree $K \in \mathbb{N}$ and level $L \in \mathbb{N}$, if there are functions $h_{lj} \colon \mathbb{R}^K \to \mathbb{R}$, $1 \le l \le L, 1 \le j \le n$, such that

$$g(x) = \sum_{j=1}^n \beta_j z_j^L,$$

where

$$z_j^l = h_{lj}(\alpha_{lj1} z_{i_{lj1}}^{l-1}, \dots, \alpha_{ljK} z_{i_{ljK}}^{l-1}),$$

for $l = 1, \dots, L$, $z_i^0 = x_i$, $i_{ljk} \in \{1, \dots, n\}$ and α_{ljk}, $\beta_j \in \mathbb{R}$.

Let us once again consider the setting of Theorem 2 and define $\Omega_{K,L}^n$ as the set of compositional functions with degree K and level L such that $h_{lj} \in W_r^n$, $1 \le l \le L$, $1 \le j \le n$. Then we can construct a neural network with L hidden layers, where each activation function is a $\mathcal{C}^\infty$-function and not a polynomial, that is able to represent functions in $\Omega_{K,L}^n$ efficiently. More precisely, under certain regularity assumptions on the functions approximated in the single layers, it can be shown that for all $g \in \Omega_{K,L}^n$ it holds that

$$\inf_\theta \|W(x; \theta) - g(x)\|_{\infty, K_n} \le \epsilon,$$

with a number N_ϵ of neurons that satisfies

$$N_\epsilon = \mathcal{O}(n^{1 + \frac{K}{r}} \epsilon^{-\frac{K}{r}}). \tag{4}$$

In total, we observe that for compositional functions with fixed degree K and level L, the required number of neurons grows only polynomially in the state dimension n, whence the curse of dimensionality is avoided.

Thus, in this talk, we discuss conditions regarding the system (1) such that this result can be applied for approximating a control Lyapunov function. To this end, we focus on ensuring compositionality of control Lyapunov functions. Furthermore, we present a corresponding network architecture that enables us to overcome the curse of dimensionality for control Lyapunov functions.

4. COMPUTING LYAPUNOV FUNCTIONS USING DEEP NEURAL NETWORKS

Let us first consider the particular case, where f in (1) does not depend on u, i.e., we have an ordinary differential equation and want to compute a Lyapunov function. For this case, we briefly present the approach from Grüne (2021) that is extended in this talk. There a small-gain condition (see, e.g., Dashkovskiy et al. (2010)) has been used to establish the existence of a Lyapunov function of the form

$$V(x) = \sum_{i=1}^s \hat{V}_i(z_i), \tag{5}$$

where $\hat{V}_i \colon \mathbb{R}^{d_i} \to \mathbb{R}$ and $z_i \in \mathbb{R}^{d_i}$ such that $x = (z_1, \dots, z_s)$. Note that V is a compositional function of level 1 and degree $K = \max_{1 \le i \le s} d_i$.

Let $K \in \mathbb{N}$ and $c > 0$ be fixed and define $\mathcal{F}_K$ as the set of Lipschitz functions $f \colon \mathbb{R}^n \to \mathbb{R}^n$, $n \in \mathbb{N}$, such that the ordinary differential equation $\dot{x}(t) = f(x(t))$ possesses a Lyapunov function of the form (5) with $\max_{1 \le i \le s} d_i \le K$. We can then construct a neural network that approximates

for every $f \in \mathcal{F}_K$ such a Lyapunov function and has a complexity of

$$N_\epsilon = \mathcal{O}(n^{1+K}\epsilon^{-K}),$$

see (4). However, for the construction of the neural network, the separation of the state space into subsystems $x = (z_1, \dots, z_s)$ would have to be known beforehand. It is shown in Grüne (2021) that this can be circumvented by appending an additional layer to the neural network that represents a linear transformation of the original system into the desired form.

For the approximation of a Lyapunov function, our network has to satisfy the conditions (2) and (3) adapted to the uncontrolled case, i.e.,

$$\begin{aligned}&\alpha_1(\|x\|) \le W(x;\theta) \le \alpha_2(\|x\|),\\ &\inf_{u\in U} DW(x;\theta)f(x) \le -\alpha_3(\|x\|) < 0.\end{aligned}$$

To this end, we formulate a cost function L that depends on the output $W(x,\theta)$ as well as on its derivative in direction of the vector field $DW(x,\theta)f(x)$:

$$\begin{aligned}&L(x, W(x;\theta), DW(x;\theta)f(x))\\ &:= ([W(x;\theta) - \alpha_1(\|x\|)]_-)^2 + ([W(x;\theta) - \alpha_2(\|x\|)]_+)^2\\ &\quad +\mu\left([DW(x;\theta)f(x) + \alpha_3(\|x\|)]_+\right)^2,\end{aligned}$$

where $[\cdot]_+ := \max(\cdot, 0)$, $[\cdot]_- := \min(\cdot, 0)$ and $\mu > 0$ is a weighting factor. Using this cost function, the corresponding neural network can be trained towards a Lyapunov function via reinforcement learning.

5. COMPUTING CONTROL LYAPUNOV FUNCTIONS

In order to extend the approach discussed in Section 4 to the controlled case, we investigate structural properties for (1) that yield the existence of a compositional control Lyapunov function. On the other hand, we also construct a suitable network architecture and cost function. In particular, we deal with the appearance of the infimum in our cost function through equation (3).

For establishing compositionality of control Lyapunov functions, we discuss the use of methods from nonlinear control theory. These methods serve the purpose of decoupling problems by decomposing them into smaller subproblems. In our context, such a decomposition into subproblems can be used to obtain a compositional form of the respective control Lyapunov function.

Consider a system of the form

$$\dot{z} = f(z, \xi), \tag{6}$$

$$\dot{\xi} = g(z, \xi) + u \tag{7}$$

and assume that V_0 is a control Lyapunov function for (6) with input ξ and that γ is a stabilizing $\mathcal{C}^2$-feedback for the same system. Then the backstepping procedure (cf. Section 6.1 in Sepulchre et al. (1997)) yields that

$$V(z, \xi) := V_0(z) + \frac{1}{2}\|\xi - \gamma(z)\|^2$$

is a control Lyapunov function for the whole system (6) - (7). In this talk, we discuss how an iterative application of backstepping for a lower triangular system yields a compositional control Lyapunov function. Furthermore, we construct a neural network that approximates this function. Similar to Section 4, we cannot only handle systems that are already given in the desired lower triangular structure, but allow for all systems that have this form after a suitable linear transformation that is learned by the deep neural network.

Next to the network architecture, we also have to formulate a suitable cost function for our purpose of reinforcement learning. To this end, we distinguish two cases. Firstly, we consider the case where the expression

$$\inf_{u\in U} DW(x;\theta)f(x,u) \tag{8}$$

can be evaluated directly. This enables us to implement condition (3) into the cost function analogously to Section 4. For example, if the system (1) is of the form

$$\dot{x}(t) = f(x(t), u(t)) = h(x(t)) + g(x(t))u(t)$$

with $U = [-c, c]^m$ for some $c > 0$, we have

$$\begin{aligned}\inf_{u\in U} DW(x;\theta)f(x,u) =& DW(x;\theta)h(x)\\ &-c\|DW(x;\theta)g(x)\|_1.\end{aligned}$$

Secondly, for the cases where such an explicit calculation of the expression (8) is not possible, we discuss methods for learning an approximation of the respective control values alongside the control Lyapunov function. This also leads to the question, under which conditions a compositional control Lyapunov function implies the existence of a compositional stabilizing feedback.

Further, in this talk we demonstrate how these and similar approaches perform in practice.

REFERENCES

Darbon, J., Langlois, G., and Meng, T. (2020). Overcoming the curse of dimensionality for some Hamilton-Jacobi partial differential equations via neural network architectures. *Research in the Mathematical Sciences*, 7. doi:10.1007/s40687-020-00215-6.

Dashkovskiy, S., Rüffer, B., and Wirth, F. (2010). Small gain theorems for large scale systems and construction of ISS Lyapunov functions. *SIAM Journal on Control and Optimization*, 48(6), 4089–4118.

Gaby, N., Zhang, F., and Ye, X. (2021). Lyapunov-Net: A deep neural network architecture for Lyapunov function approximation. Preprint. ArXiv 2109.13359.

Grüne, L. (2021). Computing Lyapunov functions using deep neural networks. *Journal of Computational Dynamics*, 8(2), 131–152. doi:10.3934/jcd.2021006.

Han, J., Jentzen, A., and E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34), 8505–8510. doi: 10.1073/pnas.1718942115.

Kang, W., Sun, K., and Xu, L. (2021). Data-driven computational methods for the domain of attraction and Zubov's equation. Preprint. ArXiv 2112.14415.

Khansari-Zadeh, S. and Billard, A. (2014). Learning control Lyapunov function to ensure stability of dynamical system-based robot reaching motions. *Robotics and Autonomous Systems*, 62(6), 752–765.

Ledyaev, Y. and Sontag, E. (1999). A Lyapunov characterization of robust stabilization. *Nonlinear Analysis: Theory, Methods & Applications*, 37, 813–840.

Long, Y. and Bayoumi, M. (1993). Feedback stabilization: Control Lyapunov functions modelled by neural

networks. In *Proceedings of 32nd IEEE Conference on Decision and Control*, 2812–2814. IEEE.

Mhaskar, H. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8(1), 164–177. doi:10.1162/neco.1996.8.1.164.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5), 503–519. doi:10.1007/s11633-017-1054-2.

Richards, S., Berkenkamp, F., and Krause, A. (2018). The Lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. In *Proceedings of The 2nd Conference on Robot Learning*, Proceedings of Machine Learning Research, 466–476. PMLR.

Sepulchre, R., Janković, M., and Kokotović, P. (1997). *Constructive Nonlinear Control*. Springer, London.

Sontag, E. (1983). A Lyapunov-like characterization of asymptotic controllability. *SIAM Journal on Control and Optimization*, 21(3), 462–471. doi:10.1137/0321028.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Reinforced Likelihood Box Particle Filter

Quoc Hung LU * **Soheib FERGANI** * **Carine JAUBERTHIE** *

* *LAAS-CNRS, Université de Toulouse, CNRS, UPS, Toulouse, France*
e-mail: qhlu@laas.fr, sfergani@laas.fr, cjaubert@laas.fr

Abstract: This paper is concerned with the development of a general scheme for box particle filtering, in which the likelihood computation is shown to be the most crucial step for the estimation strategy. An overview on Box Particle Filters and discussions about from assumptions used in the literature to the filters performance evaluation approach are in the scope of the paper. From this, we aim to produce a filter taking advantages from strong aspects of various existing box particle filters. A class of nonlinear L^2 functions is concerned. Also, a comparative study via an illustration example to highlight the efficiency of the proposed method is investigated.

Keywords: State filtering and Estimation, Nonlinear System, Box Particle Filter, Interval Analysis; 93E10, 93E11.

1. INTRODUCTION

In State Estimation or Filtering problems, when dealing with a linear Gaussian state-space model, analytical expressions computing the state estimates according to posterior distributions can be derived by the well known and widespread Standard Kalman Filter (SKF) (Kalman, 1960). Many extension of SKF are then provided by numerous researches in different contexts (Mohamed and Nahavandi, 2012, Combastel, 2015, Chen et al., 1997, Lu et al., 2019). For nonlinear model without Gaussian measurement assumption, Particle Filters (PF) have been applied successfully to a variety of state estimation problems (Gordon et al., 1993, Doucet et al., 2001). The PF efficiency and accuracy depend mostly on the number of particles used in the estimation which may require a large computation time.

One of the famous extensions of PF to set membership approach is the Box Particle Filter (BPF) (Abdallah et al., 2008). BPF handles box (interval vector of) states and bounded errors by using interval computation and constraint satisfaction techniques. This method has been shown to control quite efficiently the number of required particles, hence reducing the computational cost and providing good results in several experiments.

Since then, numerous variants of BPF are developed (Nassreddine et al., 2010, Blesa et al., 2015, Tran et al., 2018) to deal with measurements bounded uncertainty, measurements stochastic uncertainty or measurements mixed uncertainty. Various techniques and theories have been proposed to address the diversity of requirements in these contexts, e.g. weight updating using Bayesian filtering technique extending to box particle case (Blesa et al., 2015) or belief function theory with different methods (Nassreddine et al., 2010, Tran et al., 2018).

In the present work, regarding this large variety of BPF, a scheme is proposed to give a generalized description that highlights the specificity of this class of filters. An analysis of the likelihood computation methodology is investigated. Furthermore, the system under consideration is nonlinear and concerned a concrete class of functions which is the L^2 space. Throughout these developments, we aim to produce a novel filter benefiting the advantages of existing methods with a number of reinforcement techniques.

2. PROBLEM FORMULATION

2.1 Notations and definitions

A real interval matrix $[X]$ of dimension $p \times q$ is a matrix with real interval components $[x_{ij}]$, $i \in \{1, ..., p\}$, $j \in \{1, ..., q\}$. Write $X \in [X]$ to indicate a point matrix $X = (x_{ij})$ belonging element-wise to $[X]$. Define:

$$\overline{X} \equiv \sup([X]) \triangleq (\sup([x_{ij}])),$$
$$\underline{X} \equiv \inf([X]) \triangleq (\inf([x_{ij}])),$$

as element-wise operators applying to $[X]$ and then $\text{mid}([X]) \triangleq (\overline{X} + \underline{X})/2$, $\text{rad}([X]) \triangleq (\overline{X} - \underline{X})/2$, $\text{width}([X]) \triangleq \overline{X} - \underline{X}$. Define also the (convex) hull of two interval matrices $[X_1]$, $[X_2]$ of the same dimension as $\text{hull}\{[X_1], [X_2]\} \triangleq [\min\{\underline{X}_1, \underline{X}_2\}, \max\{\overline{X}_1, \overline{X}_2\}]$.

Basic interval operators $\diamond \in \{+, -, \times, \div\}$ defined in Jaulin et al. (2001) can be used to compute directly all operations $[u] \diamond [v]$ and $\alpha \diamond [u]$, for real intervals $[u]$, $[v]$ and $\alpha \in \mathbb{R}$, without any further approximation algorithm. Then, interval matrix computations are defined similarly to matrix computations using the basic operators while more general operators are constructed by meant of inclusion function $[f]$ (Jaulin et al., 2001). In practice, the package Intlab developed for Matlab is used for computations.

2.2 Assumptions and discussions

Consider the following dynamical system:

$$(\Sigma): \quad \begin{cases} x_k = f(x_{k-1}, u_k, w_k) \\ y_k = h(x_k, u_k, v_k) \end{cases}, \quad k \in \mathbb{N}^*, \tag{1}$$

where $x_k \in \mathbb{R}^{n_x}$ and $y_k \in \mathbb{R}^{n_y}$ are respectively state and output measurement, $u_k \in \mathbb{R}^{n_u}$ input, $w_k \in \mathbb{R}^{n_w}$ state dynamic disturbance and $v_k \in \mathbb{R}^{n_y}$ measurement noise.

***Assumption (A)*: State Process Uncertainty**
u_k and w_k are unknown and belong to known intervals $[u_k]$ and $[w_k]$ respectively.

***Assumption (B)*: Measurement Bounded Uncertainty**
(B1) v_k (unknown) belongs to known interval $[v_k]$.
(B2) The measurements are intervals $[y_k]$.
(B3) The measurements are assumed to be accurate in the sense that $[y_k] \ni h(x_k, u_k) \equiv h(x_k, u_k, 0)$ (the zero noise case), where x_k is the real state .

***Assumption (C)*: Measurement Stochastic Uncertainty**
(C1) v_k are additive noises with known density p_v.
(C2) The measurements are point values y_k.

***Assumption (D)*: Measurement Mixed Uncertainty**
(D1) v_k are additive Gaussian noises with unknown mean $\mu_k \in \mathbb{R}^{n_y}$ and covariance $\Sigma_k \in \mathbb{R}^{n_y \times n_y}$.
(D2) μ_k, Σ_k belong to known intervals $[\mu_k]$, $[\Sigma_k]$.
(D3) The measurements are point values y_k.

Assumption (A) is used in Abdallah et al. (2008), Nassreddine et al. (2010), Blesa et al. (2015), Tran et al. (2018).

Assumptions (B) are under study in Abdallah et al. (2008), Nassreddine et al. (2010). In Abdallah et al. (2008), the BPF is introduced and becomes standard for many extensions or variants with essential steps: *initialization, box propagation, contraction, likelihood (weight) computation, state estimation and resampling*. In Nassreddine et al. (2010), the Belief State Estimation algorithm is developed using the belief function theory. It may require some techniques for the construction and computation of masses, but after being normalized, these masses become likelihoods in the probability sense. Therefore, we also call likelihood computation as an essential step of this method.

Assumptions (C) are used in Blesa et al. (2015). The method proposed therein includes a different approach to weight the box particles as well as a resampling procedure based on repartitioning the box enclosing the propagated states. There is no contraction step in this method.

Assumptions (D) are used in Tran et al. (2018), in which (D1) is a special case of (C1) with a slight relaxation by adding bounded uncertainties to Gaussian parameters μ_k and Σ_k. In Tran et al. (2018), the belief function theory is used with continuous mass functions to represent these kinds of uncertainties and to compute box particle likelihoods. The proposed approach therein leads to the so-called Evidential Box Particle Filter (EBPF) including all essential steps of the standard BPF.

Remark 1. (B3) is the implicit assumption deriving the consistency between the predicted measurement boxes $[h]([x_k^i], [u_k])$, $i \in \{1, ..., M\}$ (M the number of partitioned boxes), and the real measurement box $[y_k]$. This consistency is used in the contraction step and the likelihood computation by penalizing all particle boxes with which the intersections $[h]([x_k^i], [u_k]) \cap [y_k]$ are empty.

Remark 2. Assumptions (D3) and (C2) are coincided. They can be transformed into (B2) with a slight relaxation of (B3). That is, knowing the density of v_k, we deduce its confidence intervals $[v_k]$ with some significant level α and define $[y_k] \triangleq y_k - [v_k]$. Then (B3) is relaxed in the sense that the observed measurements $[y_k]$ do not contain $h(x_k, u_k)$ with certainty but with only a high probability $(1 - \alpha)$.

3. GENERAL SCHEME OF BOX PARTICLE FILTER

3.1 Scheme

Although applying different background theories, the proposed methods in Abdallah et al. (2008), Nassreddine et al. (2010), Blesa et al. (2015), Tran et al. (2018) study State Estimation in a framework of stochastic uncertainties and/or bounded uncertainties with two main objectives:

Objective 1: Reduce as much as possible the width of box particles to penalize the conservatism due to interval computations (the wrapping effect).

Objective 2: Quantify (compute) box particle likelihoods as well as possible to enhance the accuracy of the estimates.

The methods used in these references can be considered as variants of BPF and be summarized by Scheme 1 which is applied in a mostly similar manner across them.

Remark 3. In this Scheme, for a general presentation, the observed measurements are denoted as intervals since the point values are considered as special cases of intervals. N_{k_0} in the initialization step takes value in $\{1, ..., M\}$ and is the number of box particles obtained at the end of the likelihood computation step at the previous time instant $(k_0 - 1)$. For $k_0 = 0$, the initialization concerns only the partition of $[x_0]$ and not the resampling. The Condition C in the while loop is different from method to method.

Remark 4. BPFs often use a non large (small) number of particles to gain computation time and reduce the loss of a guaranteed estimation. Consequently, the resampling or repartition step happens almost always, at every or only after a few iterations. Therefore, the fact that we hold previous weights and update them afterward has no significant effect while this effect might not be quantified easily. Furthermore, conditions under which the resampling or repartition is implemented base usually on some heuristic choice of a threshold. This is also an issue of discussion.

3.2 Performance evaluation of BPFs sharing Scheme 1

In order to evaluate how the computed likelihoods bring efficiency to the estimation, it must compare the result of a BPF with that of the basic scenarios of Scheme 1:

- **Scenario 1**: Using the contraction step without partition (1 box particle);
- **Scenario 2**: Using equi-likelihood $1/M$ and without contraction step (M box particles);
- **Scenario 3**: Using equi-likelihood $1/M$ and with contraction step (M box particles).

The reason is that, in some applications, using solely the contraction step, the algorithm performance has been rather good and the efficiency brought by the computed

Scheme 1 General Scheme of Box Particle Filtering

STEP 1. Initialization. At a time step $k_0 \geq 0$, (re)partition the interval $[x_{k_0}]$ or resample the set $\{[x^j_{k_0}], w_j\}_{j=1:N_{k_0}}$ into M disjoints equal-volume sub-boxes with the same weights: $\{[x^i_{k_0}], w_i = 1/M\}_{i=1:M}$.

while $\{[x^i_{k_0}], w_i\}_{i=1,...,M}$ still satisfies a predetermined Condition C **do**

STEP 2. Propagation. Get a new set of box particles $\{[x^i_{k_0+1}] = [f]([x^i_{k_0}], [u_{k_0}])\}_{i=1,...,M}$ estimating the box containing the real state $x_{k_0+1} = f(x_{k_0}, u_{k_0})$ with or without a contraction step.

STEP 3. Likelihood computation

a) Compute (and normalize) the likelihoods of box particles $\{[x^i_{k_0+1}]\}_{i=1,...,M}$ being the box containing the real state x_{k_0+1}. This computation bases on the consistency between the estimated measurement $[h]([x^i_{k_0+1}], [u_{k_0}])$'s and the obtained measurement $[y_{k_0+1}]$ using different criteria and methods.

By this step, the following set of box particles with updated weights is obtained : $\{[x^i_{k_0+1}], w_i\}_{i=1,...,M}$.

b) Some techniques can be applied at this step to get a more "efficient" set of box particles, e.g. discarding the boxes with small weights (smaller than some predetermined threshold) and with or without replicating the box associated with the greatest weight,... From this, the set of box particles becomes $\{[x^i_{k_0+1}], w_i\}_{i=1,...,N_{k_0+1}}$, $1 \leq N_{k_0+1} \leq M$.

STEP 4 : Estimation

Interval estimate:

$$[x_{k_0+1}] = \sum_{i=1}^{M} w_i.[x^i_{k_0+1}] \tag{2}$$

Point estimate:

$$x_{k_0+1} = \sum_{i=1}^{M} w_i.\text{mid}([x^i_{k_0+1}]) \tag{3}$$

STEP 5 : $k_0 = k_0 + 1$

end while

STEP 6 : Restarting at STEP 1.

likelihoods might be insignificant. The same manner might happen for the other scenarios.

Following indexes, proposed in Tran et al. (2018), will be used for performance evaluations:

$$\overline{RMSE}_j = \sup \sqrt{\sum_{k=1}^{N} (x_{k,j} - [\hat{x}_{k,j}])^2 / N},$$

$$E_j = \sum_{k=1}^{N} \text{width}([\hat{x}_{k,j}])/N,$$

$$O_j = \left(\sum_{k=1}^{N} \mathbf{1}(x_{k,j} \in [\hat{x}_{k,j}])/N \right) \times 100,$$

where $j \in \{1, ..., n_x\}$, $\overline{RMSE}$ is the root mean squared error upper bound, $\mathbf{1}(x)$ equal 1 (element-wise) if the condition x holds true (element-wise) and vanishes otherwise.

3.3 Likelihood computation methodology

In the next, the diagram in Fig.1 is used to discuss the methodology of Likelihood Computation Methods (LCM) using in Scheme 1.

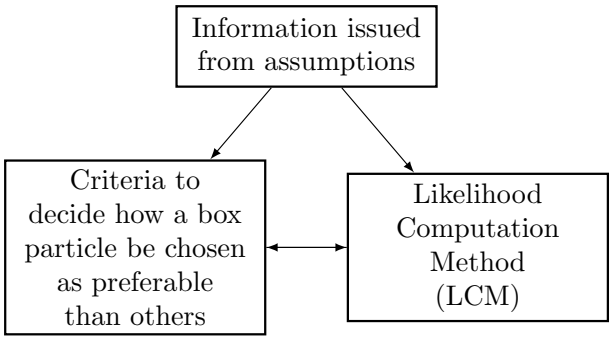


Fig. 1. Likelihood computation methodology diagram

First of all, that is the assumptions of the system under consideration supply the information needed to build the likelihood. For instance, the information may be:

- **Information (a)**: The intersection between $[y_k]$ and the box $[h]([x^i_k], [u_k])$ containing the real value y_k must be non empty,
- **Information (b)**: The distribution of v_k and hence of $r_k = y_k - h(x_k, u_k)$ is Gaussian (for additive measurement noise),
- (or more other piece of information)...

The information can be directly an assumption or a deduction of the later. In bounded-error context, only **Information (a)** is treated (Abdallah et al., 2008) while in the mixed uncertainty case, both **Information (a)** and **Information (b)** are taken into account (Tran et al., 2018).

Criteria and methods are then chosen to exploit the information. On the one hand, once a criterion is chosen, different methods can be used to calculate the likelihood. On the other hand, a calculation method may correspond to one or many criteria. There are also calculation methods that exploit better the supplied information than others.

4. CONCLUSION

Thanks to the initialized investigations aforementioned, we go to the essential of this class of BPF methods which helps to produce an enhancing method. Furthermore, the additional assumption that f and h of (1) belonging to L^2 space enlarges the problem with many theoretical and applied aspects to be solved.

REFERENCES

Abdallah, F., Gning, A., and Bonnifait, P. (2008). Box particle filtering for nonlinear state estimation using interval analysis. *Automatica*, 44(3), 807–815. doi: 10.1016/j.automatica.2007.07.024.

Blesa, J., Le Gall, F., Jauberthie, C., and Travé-Massuyès, L. (2015). State estimation and fault detection using box particle filtering with stochastic measurements. In *26th International Workshop on Principles of Diagnosis (DX-15)*, 67–73. Paris, France.

Chen, G., Wang, J., and Shieh, S. (1997). Interval Kalman filtering. *IEEE Transactions on Aerospace and Electronic Systems*, 33(1), 250–259. doi:10.1109/7.570759.

Combastel, C. (2015). Merging kalman filtering and zonotopic state bounding for robust fault detection under noisy environment. *IFAC-PapersOnLine*, 48(21), 289–295. doi:10.1016/j.ifacol.2015.09.542. 9th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2015.

Doucet, A., de Freitas, N., and Gordon, N. (eds.) (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Gordon, N.J., Salmond, D.J., and Smith, A.F.M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2), 107–113. doi:10.1049/ip-f-2.1993.0015.

Jaulin, L., Kieffer, M., Didrit, O., and Walter, E. (2001). *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics*. Springer-Verlag, London.

Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1), 35–45. doi:10.1115/1.3662552.

Lu, Q.H., Fergani, S., Jauberthie, C., and Le Gall, F. (2019). Optimally bounded interval kalman filter. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 379–384. doi:10.1109/CDC40024.2019.9028918.

Mohamed, S.M.K. and Nahavandi, S. (2012). Robust finite-horizon kalman filtering for uncertain discrete-time systems. *IEEE Transactions on Automatic Control*, 57(6), 1548–1552. doi:10.1109/TAC.2011.2174697.

Nassreddine, G., Abdallah, F., and Denoux, T. (2010). State estimation using interval analysis and belief-function theory: Application to dynamic vehicle localization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(5), 1205–1218. doi:10.1109/TSMCB.2009.2035707.

Tran, T.A., Jauberthie, C., Le Gall, F., and Travé-Massuyès, L. (2018). Evidential box particle filter using belief function theory. *International Journal of Approximate Reasoning*, 93, 40–58. doi:10.1016/j.ijar.2017.10.028.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

An untold story in system identification: a necessary and sufficient condition for system identifiability from finite data

M.K. Çamlibel * **P. Rapisarda** **

* *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands, (e-mail:* `m.k.camlibel@rug.nl`*)*
** *School of Electronics and Computer Science, University of Southampton, United Kingdom (e-mail:* `pr3@ecs.soton.ac.uk`*)*

Abstract: We state necessary and sufficient conditions for one finite length input-output trajectory to determine uniquely (modulo state-space isomorphisms) a minimal linear, deterministic input-state-output system, given an upper bound on the state dimension.

Keywords: Time-series modelling, time-invariant systems, nonparametric methods

1. INTRODUCTION

Given a *minimal* linear input-state-output (i-s-o) system

$$\begin{aligned} x(t+1) &= A_{\text{true}}x(t) + B_{\text{true}}u(t) \\ y(t) &= C_{\text{true}}x(t) + D_{\text{true}}u(t)\ , \end{aligned} \tag{1}$$

where $A_{\text{true}} \in \mathbb{R}^{n_{\text{true}}\times n_{\text{true}}}$, $B_{\text{true}} \in \mathbb{R}^{n_{\text{true}}\times m}$, $C_{\text{true}} \in \mathbb{R}^{p\times n_{\text{true}}}$ and $D_{\text{true}} \in \mathbb{R}^{p\times m}$ and $T \geqslant 0$, denote by

$$\begin{aligned} u_{[0,T]} &:= [u(0)\ \dots\ u(T)] \in \mathbb{R}^{m\times(T+1)} \\ y_{[0,T]} &:= [y(0)\ \dots\ y(T)] \in \mathbb{R}^{p\times(T+1)}\ , \end{aligned} \tag{2}$$

a finite input-output (i-o) trajectory generated by (1). The *deterministic system identifiability* problem is usually stated as follows:

> *Determine (necessary and) sufficient conditions on $u_{[0,T]}$ and $y_{[0,T]}$ such that (1) can be uniquely identified, up to a nonsingular transformation of the state variable.*

Standard references on deterministic system identifiability with non-impulsive inputs are Grewal and Glover (1976), Sontag (1980), Kalman (1983), Gevers and Wertz (1984), Willems et al. (2005). More recent publications dealing with finite length data are Heij (1993), Markovsky et al. (2005), Markovsky and Dörfler (2020). The relation of some of these results with ours is deferred to remarks interspersed in the text.

We study identifiability *relative to a model class*. A priori knowledge on (1), or properties required of a model compatible with the data, are integrated in the problem formulation by specifying the *model class* $\mathcal{S}_{pk}$ that the identified systems belong to. $\mathcal{S}_{pk}$ consists e.g. of systems with a given upper bound on the state-space dimension; of controllable systems; of minimal ones; and so on. We use a *data informativity* (see van Waarde et al. (2020)) perspective: an i-o trajectory is informative about a property (e.g. stability, controllability) if it is shared by *all* systems compatible with the data. The problem of *informativity for system identification* is:

> *Given $\mathcal{S}_{pk}$ establish necessary and sufficient conditions on $u_{[0,T]}$ and $y_{[0,T]}$ such that a model in $\mathcal{S}_{pk}$ compatible with the data can be uniquely identified, up to a nonsingular transformation of the state variable.*

The deterministic system identifiability problem is a special case of this one. Due to space limitations, we only state *necessary and sufficient* conditions for identifiability when $\mathcal{S}_{pk}$ consists of *minimal i-s-o systems with a given upper bound on the state dimension*, and we omit the proofs of the necessary results. A complete illustration of our results will be given elsewhere.

Notation and terminology

Intervals. Given $i, j \in \mathbb{N}\cup\{0\}$, $i \leqslant j$, we denote by $[i, j]$ the set $[i, j] := \{k \in \mathbb{N}\cup\{0\} \mid i \leqslant k \leqslant j\}$.

Matrices and vectors. The set of $n \times m$ matrices with real entries is denoted by $\mathbb{R}^{n\times m}$. If one of the dimensions of $M \in \mathbb{R}^{n\times m}$ is zero, then we take $\mathbb{R}^{n\times m}$ to denote the empty set, and M to denote a void matrix, i.e. a matrix with zero rows and zero columns.

Hankel matrices. Given $T \in \mathbb{N}\cup\{0\}$, a set of m-dimensional vectors $\{f_t\}_{t=0,\dots,T}$, and $0 \leqslant i \leqslant T$, we denote by

$$f_{[i,T]} := [f_i\ \dots\ f_T] \in \mathbb{R}^{m\times(T-i+1)},$$

the *matrix associated with* $\{f_t\}_{t=0,\dots,T}$.

Given $f_{[i,j]} = [f_i\ \cdots\ f_j] \in \mathbb{R}^{m\times(j-i+1)}$ and k with $j - k \in [i, j]$, the *Hankel matrix of $f_{[i,j]}$ with depth $k+1$* is the $(k + 1)m \times (j - k - i + 1)$ matrix defined by:

$$H_k(f_{[i,j]}) = \begin{bmatrix} f_i & \cdots & f_{j-k} \\ \vdots & & \vdots \\ f_{i+k} & \cdots & f_j \end{bmatrix}\ . \tag{3}$$

The "depth" is the number of block-rows of (3).

Dynamical systems. Given an i-s-o discrete-time system

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$$x(t+1) = Ax(t) + Bu(t)$$
$$y(t) = Cx(t) + Du(t)\ , \tag{4}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$, and $k \in \mathbb{N}$, we define the *k-th observability matrix* by

$$\Omega_k := \begin{bmatrix} C^\top & (CA)^\top & \dots & \left(CA^k\right)^\top \end{bmatrix}^\top \ .$$

We denote by $\ell(C, A)$ the smallest integer $k \in \mathbb{N}$ such that $\operatorname{rank}\Omega_k = \operatorname{rank}\Omega_{k-1}$. Note that $\ell(C, A) \leqslant n$; if (C, A) is observable, then $\ell(C, A)$ is the observability index of (4).

2. PROBLEM FORMULATION

The set of all systems (4) is denoted by

$$\mathcal{S} := \left\{ \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathbb{R}^{(n+p)\times(n+m)} \mid n \geqslant 0 \right\} \ . \tag{5}$$

$\mathcal{S}$ includes memoryless systems ($n = 0$), in which case A, B and C are void.

Definition 1. (Explanatory system). A system $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{S}$ with $A \in \mathbb{R}^{n\times n}$ *explains the data* $(u_{[0,T]}, y_{[0,T]})$ if there exists $x_{[0,T]} \in \mathbb{R}^{n\times(T+1)}$ such that

$$x_{[1,T]} = Ax_{[0,T-1]} + Bu_{[0,T-1]}$$
$$y_{[0,T]} = Cx_{[0,T]} + Du_{[0,T]}\ . \tag{6}$$

The set of all systems explaining $u_{[0,T]}$ and $y_{[0,T]}$ is denoted by $\mathcal{E}$ and called (the class of) *explanatory systems*:

$$\mathcal{E} := \{ \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{S} \mid \text{ there exists } n \geqslant 0 \text{ and } x_{[0,T]} \in \mathbb{R}^{n\times(T+1)} \text{ such that (6) holds} \} \ . \tag{7}$$

We are also interested in subclasses of explanatory systems such as those with a given state space dimension and those with a given lag, respectively defined by

$$\mathcal{E}(n) := \left\{ \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{E} \mid A \in \mathbb{R}^{n\times n} \right\},$$
$$\mathcal{E}(\ell, n) := \left\{ \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{E}(n) \mid \ell = \ell(C, A) \right\}.$$

The system (1) belongs to $\mathcal{E}$, $\mathcal{E}(n_{\text{true}})$, and $\mathcal{E}(\ell_{\text{true}}, n_{\text{true}})$.

It is straightforward to check that $\mathcal{E}(n)$ and $\mathcal{E}(\ell, n)$ are invariant under nonsingular state space transformations.

Definition 2. (Isomorphism property). $\begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix} \in \mathcal{E}(n)$, $i = 1, 2$ with $n \geqslant 1$ are called *isomorphic* if $D_1 = D_2$ and there exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that $A_1 = S^{-1}A_2S$, $B_1 = S^{-1}B_2$, $C_1 = C_2S$.

If $n \geqslant 1$, $\mathcal{E}(n)$ has the *isomorphism property* if *all* systems in it are isomorphic to each other. If $n = 0$, we say that $\mathcal{E}(0)$ has the *isomorphism property* if it is a singleton.

To formalize prior knowledge or assumptions about the 'true' system (1), a subset $\mathcal{S}_{\text{pk}} \subseteq \mathcal{S}$ is defined. For example, bounds on the state dimension or on the observability index are formalized defining $\mathcal{S}_{\text{pk}}$ by

$$\mathcal{S}_N := \left\{ \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{S} \mid A \in \mathbb{R}^{n\times n} \text{ with } n \leqslant N \right\}$$
$$\mathcal{S}_{L,N} := \left\{ \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{S}_N \mid \ell(C, A) \leqslant L \right\} \ . \tag{8}$$

and the class of minimal models defining $\mathcal{S}_{\text{pk}}$ by

$$\mathcal{M} := \left\{ \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathcal{S} \mid \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \text{ is minimal} \right\} \ . \tag{9}$$

The following definition formalizes the property of a finite i-o sequence that we seek to characterize in this paper.

Definition 3. (Informativity for identification). Define $\mathcal{S}$ by (5), and let $\mathcal{S}_{\text{pk}} \subseteq \mathcal{S}$. The data $(u_{[0,T]}, y_{[0,T]})$

- *uniquely determine the state dimension within* $\mathcal{S}_{\text{pk}}$ if $\mathcal{E} \cap \mathcal{S}_{\text{pk}} \subseteq \mathcal{E}(n_{\text{true}})$.
- are *informative for system identification within* $\mathcal{S}_{\text{pk}}$ if they uniquely determine the state dimension in $\mathcal{S}_{\text{pk}}$, and if $\mathcal{E} \cap \mathcal{S}_{\text{pk}}$ has the isomorphism property.

The following are two standing assumptions.

Assumption A.

$$N \in \mathbb{N} \text{ is given, such that } n_{\text{true}} \leqslant N \leqslant T \ . \tag{A}$$

Assumption B.

$$u_{[0,T]} \text{ has full row rank } . \tag{B}$$

Remark 1. It is straightforward to check that condition (B) is *necessary* for data informativity for system identification, and consequently it is not restrictive. Note also that if (B) holds, then $T \geqslant m - 1$; this provides a lower bound on the number of data points to allow system identification. We give a tighter lower bound on T further in this communication; see Theorem 2 in sect. 3.

3. MAIN RESULT

Our necessary and sufficient conditions for informativity for system identification are formulated in terms of the rank of a Hankel matrix built from the data, and of some structural integers that we now introduce.

For $k \in [0, T]$, we denote by $H_k \in \mathbb{R}^{(k+1)(m+p)\times(T-k+1)}$ the *block-Hankel matrix of depth* $k + 1$ constructed from the data $(u_{[0,T]}, y_{[0,T]})$:

$$H_k := \begin{bmatrix} H_k(u_{[0,T]}) \\ H_k(y_{[0,T]}) \end{bmatrix} = \begin{bmatrix} u_0 & \cdots & u_{T-k} \\ \vdots & & \vdots \\ u_k & \cdots & u_T \\ y_0 & \cdots & y_{T-k} \\ \vdots & & \vdots \\ y_k & \cdots & y_T \end{bmatrix} \ , \tag{10}$$

and by $G_k \in \mathbb{R}^{((k+1)m+kp)\times(T-k+1)}$ the matrix

$$G_k := \begin{bmatrix} u_0 & \cdots & u_{T-k} \\ \vdots & & \vdots \\ u_k & \cdots & u_T \\ y_0 & \cdots & y_{T-k} \\ \vdots & & \vdots \\ y_{k-1} & \cdots & y_{T-1} \end{bmatrix} \ . \tag{11}$$

From H_k and G_k we define

$$\rho_k := \begin{cases} p & \text{if } k = -1 \\ \operatorname{rank} H_k - \operatorname{rank} G_k & \text{if } k \in [0, T] \ . \end{cases} \tag{12}$$

Note that $0 \leqslant \rho_k \leqslant p$ for all k. The Hankel structure of H_k and G_k implies that $\rho_k = 0$ if and only if every annihilator of the i-o sequence with lag $\leqslant k$ is the linear combination of annihilators with lag $\leqslant k-1$ and their shifts. Equivalently, $\rho_k > 0$ if and only if there exists an annihilator of the i-o sequence with lag $\leqslant k$ that is not "implied" by lower lag annihilators and their shifts.

The following result can be proved exploiting the Hankel structure of the matrices H_k and G_k.

Lemma 1. Define s_k by

$$s_k := \rho_{k-1} - \rho_k \quad \text{for} \quad k \in [0, T] \; ; \tag{13}$$

then $s_k \geqslant 0$ for all $k \in [0, T]$. Moreover, $\sum_{i=0}^{T} s_i = p$.

Remark 2. The s_i's are related to the integers ζ_t introduced on p. 569 in Willems (1986) in the context of system identification from infinite length data, associated with the *shortest lag description* of a behavior (p. 569 *ibid.*). Moreover, the maximal lag in the shortest lag description equals the integer $\ell(A, C)$ defined in section 1 (see statement *(vii)* Theorem 6 p. 570 *ibid.*).

We denote by q the smallest integer such that $\sum_{i=0}^{q} s_i = p$:

$$q := \min \left\{ k \in [0, T] \mid \sum_{i=0}^{k} s_i = p \right\} \; . \tag{14}$$

Remark 3. Lemma 1 implies that q is well defined, and can be computed directly from the data, by computing ρ_k from (12), and s_k from (13).

We define *the minimum number of states* $n_{\min}$ and *the shortest lag* $\ell_{\min}$ by:

$$\begin{aligned} n_{\min} &:= \min\{n \geqslant 0 \mid \mathcal{E}(n) \neq \varnothing\} \\ \ell_{\min} &:= \min\{\ell \geqslant 0 \mid \exists\, n \geqslant 0 \text{ such that } \mathcal{E}(\ell, n) \neq \varnothing\} \; . \end{aligned} \tag{15}$$

The following result shows that these two integers can be computed in terms of the integers s_k and q, and consequently, in view of Lemma 1, *directly from the data*, via linear algebraic computations involving the matrices H_k and G_k defined in (10) and (11).

Theorem 1. Define $n_{\min}$ and $\ell_{\min}$ by (15), and q by (14). Then

$$\ell_{\min} = q \;\; \text{and} \;\; n_{\min} = \sum_{i=0}^{\ell_{\min}} i s_i \; .$$

Moreover,

$$\mathcal{E}(n_{\min}) = \mathcal{E}(\ell_{\text{true}}, n_{\min}) = \mathcal{E}(n_{\min}) \cap \mathcal{O} \; .$$

Remark 4. The proof of Theorem 1 is based on a series of intermediate results, some of independent interest. Most prominent among these is an iterative procedure to construct a state sequence with $n_{\min}$ components from a given finite length i-o trajectory. The procedure uses the left-annihilators of the Hankel matrices (10) of the data; it is a modification of the *shift-and-cut procedure* introduced in Rapisarda and Willems (1997). An explanatory model (4) can be straightforwardly computed from the finite length state-trajectory obtained in this way.

This algorithm offers an alternative to subspace identification procedures in the case of finite measurements, *without assumptions on the length of the data set*. On subspace identification for finite length data, see Markovsky et al. (2005), where such procedures are formulated under the assumption that the number of available measurements is at least twice the maximum lag of the system.

Remark 5. It follows from the relation between the s_k's and the shortest lag description of a behavior (see Remark 2) that the equality $n_{\min} = \sum_{i=0}^{\ell_{\min}} i s_i$ in the second claim of Theorem 1 is analogous to statement (v) of Theorem 6 of Willems (1986).

The main result of this paper is the following characterization of informativity for system identification within the class of minimal systems with an upper bound on the state dimension.

Theorem 2. Assume conditions (A) and (B). Define $n_{\min}$ and $\ell_{\min}$ by (15), $\mathcal{S}_N$ by (8), $\mathcal{M}$ by (9). Moreover, define

$$d := N - n_{\min} + \ell_{\min} \; .$$

The data $(u_{[0,T]}, y_{[0,T]})$ are informative for system identification within $\mathcal{S}_N \cap \mathcal{M}$ if and only if the following two conditions hold:

$$\begin{aligned} T + 1 &\geqslant (d+1)m + d + n_{\min} \\ \operatorname{rank} H_d &= (d+1)m + n_{\min} \; . \end{aligned} \tag{16}$$

Remark 6. The constraints (16) represent *truly data-based* necessary and sufficient conditions for system identifiability: to verify them, one only needs to know $\ell_{\min}$ and $n_{\min}$. It follows from Theorem 1 and Remark 3 that both integers are computable directly from the data.

Remark 7. The first constraint in (16) improves on the trivial lower bound $T \geqslant m - 1$ formulated in Remark 1. It implies that identifiability is possible only if a *minimal number of measurements* is available.

Remark 8. (The SISO case). When $m = p = 1$, it can be shown that $\ell_{\min} = n_{\min}$ and consequently $d = N$. The conditions (16) in this case are

$$T \geqslant 2N + n_{\min} \;\; \text{and} \;\; \operatorname{rank} H_N = N + 1 + n_{\min} \; . \tag{17}$$

The relation between T and N appearing in the first condition in (17) has an intuitive interpretation. For example, a larger uncertainty about the complexity of the generating system translates to larger values of N, and implies that more measurements should be available to uniquely identify the dynamics.

In another situation, if prior knowledge about the generating system suggests a high complexity, then a larger N should be chosen; in this case, the complex dynamics require a larger number of measurements to be uniquely identified.

Remark 9. In Sontag (1980) a bound is derived on the minimum number of measurements necessary to identify a linear, discrete-time system from i-o data. However, the underlying assumption in that work is that the system starts from the *zero initial state*, while our bound holds also for the case in which the initial conditions are nonzero. Moreover, the input is assumed to be *generic*, while our result is valid also for the case of *structured* inputs. See also Example 1 below.

Remark 10. The second condition in (16) concerns the *quality* of the data: identifiability requires that the combination of initial conditions and of the input sequence is "sufficiently rich", as reflected in the rank of the Hankel matrix.

The relation between the rank of the data Hankel matrix and the dimension of the state space is known since Kalman's work on realization from impulse response data (see Kalman et al. (1969)). Such relation, together with some "persistency of excitation" conditions on the input

sequence, lies at the foundation of several results and procedures to compute explanatory models (see for example Theorem 2 in Moonen et al. (1989), or Cor. 1 in Willems et al. (2005)). The second condition in (16) is in the same spirit, but with the fundamental difference that *persistency of excitation is not assumed*. The result of Theorem 2 can consequently be applied also to the case of structured inputs.

Example 1. (State construction with structured inputs). For zero initial conditions, $T = 5$ and $u(t) = \frac{1}{2^t}$, $t \in [0, 5]$, the SISO system ($m = p = 1$) described by

$$y(t) + 3y(t-1) + 2y(t-2) = -u(t-1) + u(t-2) \ ,$$

generates the output $y_{[0,5]} = \begin{bmatrix} 0 & -1 & \frac{7}{2} & -\frac{33}{4} & \frac{143}{8} & -\frac{593}{16} \end{bmatrix}$.

It is straightforward to verify that $\operatorname{rank} H_k = k + 2$, $\operatorname{rank} G_k = k+1$, $k = 0, 1$, $\operatorname{rank} H_k = \operatorname{rank} G_k$, $k \in [2, 5]$. It follows that $s_i = 0$ for $i \in \{0, 1, 3, 4, 5\}$, and that $s_2 = 1$. It follows from the definition (14) of q and from Theorem 1 that $q = \ell_{\min} = 2$ and $n_{\min} = 2$.

When $T = 5$, it follows from Remark 8 that the two conditions of Theorem 2 reduce to $5 \geqslant 2N + 3$ and $\operatorname{rank} H_N = N + 3$.

We now show that the data is not informative for *any* bound N on the state space dimension of an explanatory system. The inequality can only be satisfied for $N \leqslant 1$. If $N = 0$, then the second condition cannot be satisfied, since $H_0 \in \mathbb{R}^{2\times 5}$ has rank 2. If $N = 1$, the second condition is satisfied if and only if $\operatorname{rank} H_1 = 4$; however

$$\operatorname{rank} H_1 = \operatorname{rank} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \frac{1}{32} \\ 0 & -1 & \frac{7}{2} & -\frac{33}{4} & \frac{143}{8} \\ -1 & \frac{7}{2} & -\frac{33}{4} & \frac{143}{8} & -\frac{593}{16} \end{bmatrix} = 3 \ .$$

The data is not informative for system identification.

The vector $[1 \ -1 \ 0 \ -2 \ -3 \ -1]$ is a left annihilator of H_2. Using the "shift-and-cut" procedure mentioned in Remark 4, the following state trajectory is computed:

$$x_{[0,5]} = \begin{bmatrix} 0 & 1 & \frac{5}{2} & -\frac{27}{4} & \frac{133}{8} & -\frac{571}{16} \\ 0 & -1 & \frac{7}{2} & -\frac{33}{4} & \frac{143}{8} & -\frac{593}{16} \end{bmatrix} \ ,$$

corresponding to the matrices A, B, C and D defined by

$$A := \begin{bmatrix} 0 & -2 \\ 1 & -3 \end{bmatrix} \ , \ B := \begin{bmatrix} 1 \\ -1 \end{bmatrix} \ , \ C := [0 \ 1] \ , \ D := 0 \ .$$

These matrices define an i-s-o explanatory model for the given i-o trajectories. It is straightforward to verify that the system described by

$$\widetilde{A} := \begin{bmatrix} 0 & -2 \\ 1 & -3 \end{bmatrix} \ , \ \widetilde{B} := \begin{bmatrix} 3 \\ \frac{2}{-2} \end{bmatrix} \ , \ \widetilde{C} := C \ , \ \widetilde{D} := D \ ,$$

is also a minimal explanatory system, with the state sequence

$$\widetilde{x}_{[0,5]} := \begin{bmatrix} 1 & \frac{3}{2} & \frac{11}{4} & -\frac{53}{8} & \frac{267}{16} & -\frac{1141}{32} \\ 0 & -1 & \frac{7}{2} & -\frac{33}{4} & \frac{143}{8} & -\frac{593}{16} \end{bmatrix} \ .$$

This system is not isomorphic to $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$: the sequences $\widetilde{x}_{[0,5]}$ and $x_{[0,5]}$ cannot be transformed into each other by a nonsingular transformation. Note also that

$$\widetilde{C} \left(zI - \widetilde{A} \right)^{-1} \widetilde{B} + \widetilde{D} = \frac{-4z + 3}{2\left(z^2 + 3z + 2\right)} \ .$$

4. CONCLUSIONS

We introduced the concept of informativity within a model class, and we gave a characterization of such property for the case of minimal models with a given upper bound on the McMillan degree.

REFERENCES

Gevers, M. and Wertz, V. (1984). Uniquely identifiable state-space and ARMA parametrizations for multivariable linear systems. *Automatica*, 20(3), 333–347.

Grewal, M. and Glover, K. (1976). Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions on Automatic Control*, 21(6), 833–837.

Heij, C. (1993). System identifiability from finite time series. *Automatica*, 29(4), 1065–1077.

Kalman, R.E. (1983). Chapter 2 - identifiability and modeling in econometrics. volume 4 of *Developments in Statistics*, 97–136. Elsevier.

Kalman, R.E., Falb, P.L., and Arbib, M.A. (1969). *Topics in Mathematical System Theory*. McGraw-Hill, New York.

Markovsky, I. and Dörfler, F. (2020). Identifiability in the behavioral setting. Available online at `http://homepages.vub.ac.be/~imarkovs/publications`.

Markovsky, I., Willems, J., and De Moor, B. (2005). State representations from finite time series. In *Proceedings of the 44th IEEE Conference on Decision and Control*, 832–835.

Moonen, M., De Moor, B., Vandenberghe, L., and Vandewalle, J. (1989). On- and off-line identification of linear state-space models. *International Journal of Control*, 49(1), 219–232.

Rapisarda, P. and Willems, J. (1997). State maps for linear systems. *SIAM Journal of Control and Optimization*, 35(3), 1053–1091.

Sontag, E. (1980). On the length of inputs necessary in order to identify a deterministic linear system. *IEEE Transactions on Automatic Control*, 25(1), 120–121.

van Waarde, H.J., Eising, J., Trentelman, H.L., and Çamlibel, M.K. (2020). Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11), 4753–4768.

Willems, J.C. (1986). From time series to linear system, Part I. Finite dimensional linear time invariant systems. *Automatica*, 22(5), 561–580.

Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B.L.M. (2005). A note on persistency of excitation. *Systems & Control Letters*, 54(4), 325–329.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A behavioral approach to data-driven control using noisy input-output data

H.J. van Waarde, M.K. Camlibel, H.L. Trentelman *
J. Eising **

* *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, the Netherlands (e-mail: h.j.van.waarde@rug.nl).*
** *Department of Mechanical and Aerospace Engineering, University of California, San Diego (e-mail: jaapeising92@gmail.com)*

Abstract: In this extended abstract we consider input-output systems described by higher order difference equations, also called autoregressive systems. We assume that we have input-output data obtained from an underlying true, but unknown, system. The problems we then consider is to determine on the basis of these data whether this unknown system is stable. We also deal with the problem of determining whether a stabilizing controller exists, and, if so, to determine one using only the data. In order to tackle these problems we heavily rely on methods from the behavioral approach to systems and control, in particular the notion of quadratic difference form.

Keywords: Data-driven control, behavioral approach, input-output data, S-lemma.

1. INTRODUCTION

A research topic that has received a lot of attention in the past few years is data-driven analysis and control. A central problem in this area is to verify certain system properties and to design control laws for an unknown dynamical system using noisy data obtained from that system. The main challenge is to do the analysis and design without the intermediate step of modeling the system using system identification, but work directly with the data instead. This has been the subject of many recent publications in the area, mainly in the context of input-output systems in state space form, see , for example, van Waarde et al. (2020); De Persis and Tesi (2020); Berberich et al. (2021); Trentelman et al. (2020).

In the present note we will leave the realm of input-output systems in state space form, and will instead work with input-output systems described by higher order difference equations, also called auto-regressive (AR) systems. We will assume that noisy input-output data have been obtained from some unknown AR system. These data are available to check stability and to verify whether a dynamic feedback controller exists that stabilizes the unknown system, and, if so, to compute such controller.

We will establish data-based tests to tackle these problems. To do this, we will heavily rely on methods from the behavioral approach to systems and control. In particular we will adopt the notion of quadratic difference form (QDF) as a framework for Lyapunov functions for autonomous systems described by higher order difference equations, see Kojima and Takaba (2005, 2006); Willems and Trentelman (1998).

We will generalize the concepts of informativity of data for quadratic stability and quadratic stabilization in the context of input-state-output systems to input-output AR systems. Our main results will be necessary and sufficient condition for informativity in terms of feasibilty of certain linear matrix inequalities (LMIs) obtained from the data. An important tool in deriving these conditions is a matrix version of Yakubovich's S-lemma, as that was recently established in van Waarde et al. (2022).

We will use the following notation. The space of all symmetric real $q \times q$ matrices is denoted by $\mathbb{S}^q$. For a given partitioned matrix

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix},$$

with $\Pi_{11} \in \mathbb{S}^q$, $\Pi_{12}^\top = \Pi_{21}$ and $\Pi_{22} \in \mathbb{S}^r$, an important role in this note will be played by the quadratic matrix inequality

$$\begin{bmatrix} I \\ Z \end{bmatrix}^\top \Pi \begin{bmatrix} I \\ Z \end{bmatrix} \geqslant 0$$

in the unknown $Z \in \mathbb{R}^{r \times q}$. The set of all solutions Z is denoted $\mathcal{Z}_r(\Pi)$. This set is nonempty and bounded if and only if $\Pi_{22} < 0$ and $\Pi_{11} - \Pi_{12}\Pi_{22}^{-1}\Pi_{21} \geqslant 0$.

2. INPUT-OUTPUT SYSTEMS IN AR-FORM AND NOISY DATA

In this note we consider input-output systems represented by auto-regressive (AR) models of the form

$$P(\sigma)\boldsymbol{y} = Q(\sigma)\boldsymbol{u}, \tag{1}$$

where σ denotes the shift operator $(\sigma \boldsymbol{f})(t) = \boldsymbol{f}(t+1)$ and $P(\xi)$ and $Q(\xi)$ are real $p \times p$ and $p \times m$ polynomial matrices of the form

$$\begin{aligned} P(\xi) &= I\xi^L + P_{L-1}\xi^{L-1} + \ldots P_1\xi + P_0, \\ Q(\xi) &= Q_{L-1}\xi^{L-1} + \ldots + Q_1\xi + Q_0. \end{aligned} \tag{2}$$

Here L is a positive integer, called the *order*. The input $\boldsymbol{u}(t)$ and output $\boldsymbol{y}(t)$ are assumed to take their values in $\mathbb{R}^m$ and $\mathbb{R}^p$, respectively. The parameters of the model are real $p \times p$ matrices $P_0, P_1 \ldots, P_{L-1}$ and $p \times m$ matrices $Q_0, Q_1, \ldots Q_{L-1}$. Note that we assume that the leading coefficient matrix of $P(\xi)$ is the $p \times p$ identity matrix. This immediately implies that $P(\xi)$ is nonsingular and that $P^{-1}(\xi)Q(\xi)$ is strictly proper. Thus, indeed, (1) represents a (strictly) causal input-output system with input $\boldsymbol{u}$ and output $\boldsymbol{y}$.

In this note, we will deal with analysis and control design for systems of the form (1), where the polynomial matrices $P(\xi)$ and $Q(\xi)$ are *unknown*. We do assume that the order L and the dimensions m and p are known. We assume that we have noisy input-output data on a given finite time interval. These data are assumed to be obtained from an underlying true (but unknown) system. In case there are no inputs, i.e. $m = 0$, we want to use these data to check whether the true system is stable. In case that control inputs are present we want to check whether there exists a stabilizing feedback controller and, if so, determine such controller using only the data. In the present section we focus on the situation that $m > 0$, i.e. control inputs are present.

As stated above, we have noisy input-output data

$$u(0), u(1), \ldots, u(T-1),\ y(0), y(1), \ldots, y(T) \tag{3}$$

on a given time interval $\{0, 1, \ldots, T\}$ with $T \geqslant L$. These noisy data are obtained from the true system. Assume that this true system is represented by (unknown) polynomial matrices $P_s(\xi)$ and $Q_s(\xi)$ of the form (2). In other words, the true system is represented by $P_s(\sigma)\boldsymbol{y} = Q_s(\sigma)\boldsymbol{u}$. We assume that the data have been obtained in the presence of unknown noise. More concretely, we assume that $u(0), u(1), \ldots, u(T)$, $y(0), y(1), \ldots, y(T)$ are samples on the interval $\{0, 1, \ldots, T\}$ of $\boldsymbol{u}$ and $\boldsymbol{y}$ that satisfy

$$P_s(\sigma)\boldsymbol{y} = Q_s(\sigma)\boldsymbol{u} + \boldsymbol{v}$$

where $\boldsymbol{v}$ is an unknown noise signal. We do put the following assumption on the noise $\boldsymbol{v}$ during the sampling interval.

Assumption 1. The noise samples $v(0), v(1), \ldots, v(T-L)$, collected in the real $p \times (T-L+1)$ matrix

$$V := [v(0)\ v(1)\ \cdots\ v(T-L)]$$

satisfy the quadratic matrix inequality

$$\begin{bmatrix} I \\ V^\top \end{bmatrix}^\top \Pi \begin{bmatrix} I \\ V^\top \end{bmatrix} \geqslant 0, \tag{4}$$

where $\Pi \in \mathbb{S}^{p+T-L+1}$ is a known partitioned matrix

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix},$$

with $\Pi_{11} \in \mathbb{S}^p$, $\Pi_{12} \in \mathbb{R}^{p \times (T-L+1)}$, $\Pi_{21} = \Pi_{12}^\top$ and $\Pi_{22} \in \mathbb{S}^{T-L+1}$. In addition, $\Pi_{22} < 0$ and the Schur complement $\Pi_{11} - \Pi_{12}\Pi_{22}^{-1}\Pi_{21} \geqslant 0$. In particular this implies that the set $\mathcal{Z}_{T-L+1}(\Pi)$ of matrices $V^\top$ that satisfy (4) is nonempty and bounded.

Now denote $q := p + m$ and denote the unknown $p \times q$ polynomial matrix $[-Q(\xi)\ P(\xi)]$ by $R(\xi)$. Also denote

$$\boldsymbol{w} := \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix}.$$

Then (1) can be written as $R(\sigma)\boldsymbol{w} = 0$. Collect the (unknown) coefficient matrices of $R(\xi)$ in the $p \times Lq$ matrix

$$R := [-Q_0\ P_0\ -Q_1\ P_1\ \cdots\ -Q_{L-1}\ P_{L-1}] \tag{5}$$

Also arrange the data $u(0), \ldots, u(T)$, $y(0), \ldots, y(T)$ into the vectors

$$w(t) = \begin{bmatrix} u(t) \\ y(t) \end{bmatrix} \quad (t = 0, 1, \ldots, T)$$

and define an associated depth L Hankel matrix by

$$H_1(w) := \begin{bmatrix} w(0) & w(1) & \cdots & w(T-L) \\ w(1) & w(2) & \cdots & w(T-L+1) \\ \vdots & \vdots & \ddots & \vdots \\ w(L-1) & w(L) & \cdots & w(T-1) \end{bmatrix}. \tag{6}$$

Furthermore, define $H_2(w) := [y(L)\ y(L+1)\ \cdots\ y(T)]$. It is then easily verified that any input-output system (1) for which the coefficient matrix R defined by (5) satisfies

$$[R \quad I] \begin{bmatrix} H_1(w) \\ H_2(w) \end{bmatrix} = V \tag{7}$$

for some V with $V^\top \in \mathcal{Z}_{T-L+1}(\Pi)$ could have generated the noisy input-output data (3). In other words, $w(0), w(1), \ldots, w(T-1), y(T)$ are samples on the interval $\{0, 1, \ldots, T\}$ of $\boldsymbol{w}$ that satisfy

$$R(\sigma)\boldsymbol{w} = \boldsymbol{v}$$

for some $\boldsymbol{v}$ satisfying Assumption 1. Therefore, if R satisfies (7) for some V with $V^\top \in \mathcal{Z}_{T-L+1}(\Pi)$, we call the system with coefficient matrix R *compatible with the data*. Now define

$$N := \begin{bmatrix} I & H_2(w) \\ 0 & H_1(w) \end{bmatrix} \Pi \begin{bmatrix} I & H_2(w) \\ 0 & H_1(w) \end{bmatrix}^\top. \tag{8}$$

Then by combining (4) and (7) we see that the system with coefficient matrix R is compatible with the data if and only if $R^\top$ satisfies the QMI

$$\begin{bmatrix} I \\ R^\top \end{bmatrix}^\top N \begin{bmatrix} I \\ R^\top \end{bmatrix} \geqslant 0, \tag{9}$$

equivalently $R^\top \in \mathcal{Z}_{Lq}(N)$. As the true system is assumed to be compatible with the given data, the set $\mathcal{Z}_{Lq}(N)$ is nonempty.

3. AUTONOMOUS AR SYSTEMS AND DATA

As already touched upon in Section 2, a special case of AR systems of the form (1) occurs if $m = 0$, i.e. the system has no inputs. In that case (1) reduces to

$$P(\sigma)\boldsymbol{y} = 0, \tag{10}$$

which, since $P(\xi)$ is a nonsingular polynomial matrix, represents an autonomous system. In this section we will briefly discuss the notion of noisy data for this special case. In fact, in this case we have only output data

$$y(0), y(1), \ldots, y(T) \tag{11}$$

on a finite time-interval $\{0, 1, \ldots, T\}$ with $T \geqslant L$. We assume that these data come from an unknown true autonomous system. Suppose this true system is represented by the unknown polynomial matrix $P_s(\xi)$, with $P_s(\xi)$ of the form (2). Again we assume that the data samples are noisy, in the sense that they are samples of a signal $\boldsymbol{y}$ that satisfies $P_s(\sigma)\boldsymbol{y} = \boldsymbol{v}$ for some $\boldsymbol{v}$ satisfying Assumption 1.

Any system in the model class of systems of the form (10) with fixed dimension p and order L is parametrized by

its coefficient matrices $P_0, P_1, \ldots, P_{L-1}$. We collect these matrices in the $p \times Lp$ matrix

$$P := [P_0 \ P_1 \ \cdots \ P_{L-1}]. \tag{12}$$

Recalling that there are no inputs present (so $w = y$), let $H(y)$ be the Hankel matrix associated with the data as given by (6), and as before partition this matrix as

$$H(y) = \begin{bmatrix} H_1(y) \\ H_2(y) \end{bmatrix},$$

where $H_1(y)$ contains the first $(L-1)p$ rows and $H_2(y)$ the last p rows. Also define

$$N := \begin{bmatrix} I & H_2(y) \\ 0 & H_1(y) \end{bmatrix} \Pi \begin{bmatrix} I & H_2(y) \\ 0 & H_1(y) \end{bmatrix}^\top. \tag{13}$$

Then as in Section 2, the autonomous system with coefficient matrices collected in the matrix P is compatible with the data if and only if

$$\begin{bmatrix} I \\ P^\top \end{bmatrix}^\top N \begin{bmatrix} I \\ P^\top \end{bmatrix} \geqslant 0, \tag{14}$$

equivalently $P^\top \in \mathcal{Z}_{Lp}(N)$.

4. QUADRATIC DIFFERENCE FORMS AND LYAPUNOV FUNCTIONS

Here we will review the basics of quadratic difference forms. Assume that N and q are positive integers and let $\Phi \in \mathbb{S}^{(N+1)q}$ be a partitioned matrix given by

$$\Phi := \begin{bmatrix} \Phi_{0,0} & \Phi_{0,1} & \cdots & \Phi_{0,N} \\ \Phi_{1,0} & \Phi_{1,1} & \cdots & \Phi_{1,N} \\ \vdots & & \ddots & \vdots \\ \Phi_{N,0} & \Phi_{N,1} & \cdots & \Phi_{N,N} \end{bmatrix}$$

with $\Phi_{i,i} \in \mathbb{S}^q$ and $\Phi_{i,j} = \Phi_{j,i}^\top$. This real symmetric matrix defines a *quadratic difference form* (QDF), to be denoted by Q_Φ. This quadratic difference form is the operator Q_Φ that maps $\mathbb{R}^q$-valued functions $\boldsymbol{w}$ on $\mathbb{Z}_+$ to $\mathbb{R}$-valued functions $Q_\Phi(\boldsymbol{w})$ on $\mathbb{Z}_+$ defined by

$$Q_\Phi(\boldsymbol{w})(t) := \sum_{k,\ell=0}^{N} \boldsymbol{w}(t+k)^\top \Phi_{k,\ell} \, \boldsymbol{w}(t+\ell).$$

For a given QDF Q_Φ, its *rate of change* along a given $\boldsymbol{w} : \mathbb{Z}_+ \to \mathbb{R}^q$ is given by $Q_\Phi(\boldsymbol{w})(t+1) - Q_\Phi(\boldsymbol{w})(t)$. It turns out that the rate of change defines a QDF itself. Indeed, by defining the matrix $\nabla\Phi \in \mathbb{S}^{(N+2)q}$ by

$$\nabla\Phi := \begin{bmatrix} 0_q & 0 \\ 0 & \Phi \end{bmatrix} - \begin{bmatrix} \Phi & 0 \\ 0 & 0_q \end{bmatrix}, \tag{15}$$

it is easily verified that

$$Q_{\nabla\Phi}(\boldsymbol{w})(t) = Q_\Phi(\boldsymbol{w})(t+1) - Q_\Phi(\boldsymbol{w})(t)$$

for all $\boldsymbol{w} : \mathbb{Z}_+ \to \mathbb{R}^q$.

Quadratic difference forms are particularly relevant in combination with AR systems. Let $R(\xi)$ be a real $p \times q$ polynomial matrix and consider the AR system $R(\sigma)\boldsymbol{w} = 0$. Let

$$\mathcal{B}(R) := \{\boldsymbol{w} \mid R(\sigma)\boldsymbol{w} = 0\}$$

be the *behavior* of this system. The QDF Q_Φ is called nonnegative on $\mathcal{B}(R)$ if $Q_\Phi(\boldsymbol{w}) \geqslant 0$ for all $\boldsymbol{w} \in \mathcal{B}(R)$. It is called positive on $\mathcal{B}(R)$ if, in addition, $Q_\Phi(\boldsymbol{w}) = 0$ if and only if $\boldsymbol{w} = 0$. We denote this as $Q_\Phi \geqslant 0$ on $\mathcal{B}(R)$ and $Q_\Phi > 0$ on $\mathcal{B}(R)$. Likewise we define nonpositivity and negativity on $\mathcal{B}(R)$.

Stability of autonomous AR systems can be characterized in terms of QDFs. In fact, the following proposition holds.

Proposition 2. Let $P(\xi)$ be a nonsingular polynomial matrix. The corresponding autonomous system $P(\sigma)\boldsymbol{y} = 0$ is stable if and only if there exists a QDF $Q_\Psi(\boldsymbol{y})$ such that $Q_\Psi \geqslant 0$ on $\mathcal{B}(R)$ and $Q_{\nabla\Psi} < 0$ on $\mathcal{B}(R)$.

For obvious reasons, we refer to Q_Ψ as a *Lyapunov function* for the AR system $P(\sigma)\boldsymbol{y} = 0$.

5. DATA-DRIVEN STABILITY ANALYSIS OF AUTONOMOUS AR SYSTEMS

In this section we study data-based stability *analysis* for autonomous systems of the form (10). Our aim is to develop a test that determines on the basis of the output data $y(0), y(1), \ldots, y(T)$ whether our true system is stable. As we saw, the data do not determine the true system uniquely. Thus we are forced to test stability for all autonomous systems that are compatible with the data, so for all systems in $\mathcal{Z}_{Lp}(N)$, with N given by (13).

In order to proceed, we will first express the existence of a Lyapunov function Q_Ψ for the autonomous system $P(\sigma)\boldsymbol{y} = 0$ in terms of a quadratic matrix inequality. This QMI involves a symmetric matrix Ψ of dimensions $Lp \times Lp$ leading to a Lyapunov function Q_Ψ, and the matrix $P = [P_0 \ P_1 \ \cdots \ P_{L-1}]$. Indeed, we have:

Theorem 3. Let $P(\xi) = I\xi^L + P_{L-1}\xi^{L-1} + \ldots + P_1\xi + P_0$ and let $P(\sigma)\boldsymbol{y} = 0$ be the corresponding autonomous system given by (10). This system is stable if and only if there exists $\Psi \in \mathbb{S}^{Lp}$, $\Psi \geqslant 0$, such that

$$\begin{bmatrix} I \\ -P \end{bmatrix}^\top \left(\begin{bmatrix} 0_p & 0 \\ 0 & \Psi \end{bmatrix} - \begin{bmatrix} \Psi & 0 \\ 0 & 0_p \end{bmatrix} \right) \begin{bmatrix} I \\ -P \end{bmatrix} < 0. \tag{16}$$

Any such Ψ defines a Lyapunov function Q_Ψ.

Based on this, we give the following definition of informativity for quadratic stability.

Definition 4. The noisy output data $y(0), y(1), \ldots, y(T)$ are called *informative for quadratic stability* if there exists a matrix $\Psi \in \mathbb{S}^{Lp}$, $\Psi \geqslant 0$ such that the QMI (16) holds for all $P = [P_0 \ P_1 \ \cdots \ P_{L-1}]$ that satisfy the QMI (14), with N defined by (13).

Informativity for quadratic stability thus means that there exists a matrix $\Psi \in \mathbb{S}^{Lp}$ such that the QDF Q_Ψ is a Lyapunov function for all systems that are compatible with the data, i.e., all systems in $\mathcal{Z}_{Lp}(N)$ are stable with a common Lyapunov function.

Below, we will formulate a necessary and sufficient condition on the data $y(0), y(1), \ldots, y(T)$ to be informative. This condition is in the form of feasibility of a linear matrix inequality. Define a matrix Z by

$$Z := \begin{bmatrix} 0_{(L-1)p \times p} & I_{(L-1)p} \end{bmatrix}. \tag{17}$$

Then we have the following theorem:

Theorem 5. Assume that $H_1(y)$ has full row rank and that

$$\bar{N} = \begin{bmatrix} [0 \ -I_p] & 0 \\ 0 & I_{Lp} \end{bmatrix}^\top \begin{bmatrix} I & H_2(y) \\ 0 & H_1(y) \end{bmatrix} \Pi \begin{bmatrix} I & H_2(y) \\ 0 & H_1(y) \end{bmatrix}^\top \begin{bmatrix} [0 \ -I_p] & 0 \\ 0 & I_{Lp} \end{bmatrix}$$

has at least one positive eigenvalue. Then the output data $y(0), y(1), \ldots, y(T)$ are informative for quadratic stability

if and only if there exists $\Phi \in \mathbb{S}^{Lp}$, $\Phi > 0$ and a scalar $\alpha \geqslant 0$ that satisfy the LMI

$$\begin{bmatrix} \Phi - \begin{bmatrix} Z^\top & 0 \end{bmatrix}^\top \Phi \begin{bmatrix} Z^\top & 0 \end{bmatrix} & \begin{bmatrix} Z^\top & 0 \end{bmatrix}^\top \Phi \\ \Phi \begin{bmatrix} Z^\top & 0 \end{bmatrix} & -\Phi \end{bmatrix} - \alpha \bar{N} \geqslant 0. \quad (18)$$

In that case the QDF Q_Ψ with $\Psi := \Phi^{-1}$ is a Lyapunov function for all systems of the form (10) compatible with the data.

6. DATA-DRIVEN STABILIZATION OF INPUT-OUTPUT AR SYSTEMS

In this section we will discuss data-driven stabilization of input-output systems in AR-form. We will work in the setup of Section 2, with systems of the form (1), with polynomial matrices of the form (2) of given degree L.

A feedback controller for the input-output system (1) with $P(\xi)$ and $Q(\xi)$ of the form (2) will be taken to be of the form

$$G(\sigma)\boldsymbol{u} = F(\sigma)\boldsymbol{y} \quad (19)$$

with

$$G(\xi) = I\xi^L + G_{L-1}\xi^{L-1} + \ldots + G_1\xi + G_0,$$
$$F(\xi) = F_{L-1}\xi^{L-1} + \ldots + F_1\xi + F_0.$$

The leading coefficient matrix of $G(\xi)$ is assumed to be the $m \times m$ identity matrix and $G_i \in \mathbb{R}^{m\times m}$, $F_i \in \mathbb{R}^{m\times p}$ for $i = 0, 1, \ldots, L-1$. The closed loop system obtained by interconnecting the system and the controller is represented by

$$\begin{bmatrix} G(\sigma) & -F(\sigma) \\ -Q(\sigma) & P(\sigma) \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} = 0. \quad (20)$$

Note that the leading coefficient matrix of the polynomial matrix in (20) is the $q \times q$ identity matrix. Hence the controlled system is autonomous. We call the controller (19) a *stabilizing* if the controlled system (20) is stable. Now define

$$C(\xi) := \begin{bmatrix} G(\xi) & -F(\xi) \end{bmatrix}.$$

Then (20) can equivalently be written as

$$\begin{bmatrix} C(\sigma) \\ R(\sigma) \end{bmatrix} \boldsymbol{w} = 0. \quad (21)$$

Collect the coefficient matrices of $F(\xi)$ and $G(\xi)$ in the coefficient matrix C defined by

$$C := \begin{bmatrix} G_0 & -F_0 & G_1 & -F_1 & \cdots & G_{L-1} & -F_{L-1} \end{bmatrix} \quad (22)$$

and recall that the coefficient matrix of $R(\xi)$ is given by (5). Then an immediate application of Theorem 3 yields:

Lemma 6. The controlled system (21) is stable if and only if there exists $\Psi \in \mathbb{S}^{Lq}$, $\Psi \geqslant 0$, such that

$$\begin{bmatrix} I_{Lq} \\ -C \\ -R \end{bmatrix}^\top \left(\begin{bmatrix} 0_q & 0 \\ 0 & \Psi \end{bmatrix} - \begin{bmatrix} \Psi & 0 \\ 0 & 0_q \end{bmatrix} \right) \begin{bmatrix} I_{Lq} \\ -C \\ -R \end{bmatrix} < 0. \quad (23)$$

This leads to the following definition.

Definition 7. We call the input-output data $u(0), u(1), \ldots, u(T), y(0), y(1), \ldots, y(T)$ *informative for quadratic stabilization* if there exist $C \in \mathbb{R}^{m\times Lq}$ and $\Psi \in \mathbb{S}^{Lq}$, $\Psi \geqslant 0$ such that the QMI (23) holds for all R that satisfy the QMI (9), with N defined by (8).

Informativity for quadratic stabilization thus means that there exists a controller $C(\sigma)\boldsymbol{w} = 0$ (equivalently, $G(\sigma)\boldsymbol{u} = F(\sigma)\boldsymbol{y}$) and a matrix $\Psi \in \mathbb{S}^{Lq}$ such that the QDF Q_Ψ is a common Lyapunov function for all closed loop systems obtained by interconnecting the controller with an arbitrary system that is compatible with the data.

We will now state necessary and sufficient conditions for informativity for quadratic stabilization. Define the matrix Z by

$$Z := \begin{bmatrix} 0_{(L-1)q\times q} & I_{(L-1)q} \end{bmatrix}. \quad (24)$$

Let N be given by (8) and define

$$\bar{N} := \begin{bmatrix} [0 & 0 & -I_p] & 0 \\ 0 & & I_{Lq} \end{bmatrix}^\top N \begin{bmatrix} [0 & 0 & -I_p] & 0 \\ 0 & & I_{Lq} \end{bmatrix}$$

Next, consider the LMI in the unknowns D and Φ given by

$$\begin{bmatrix} \Phi & \begin{bmatrix} \Phi Z^\top & D & 0 \end{bmatrix}^\top & \begin{bmatrix} \Phi Z^\top & D & 0 \end{bmatrix}^\top \\ \begin{bmatrix} \Phi Z^\top & D & 0 \end{bmatrix} & -\Phi & 0 \\ \begin{bmatrix} \Phi Z^\top & D & 0 \end{bmatrix} & 0 & \Phi \end{bmatrix} - \alpha \begin{bmatrix} \bar{N} & 0 \\ 0 & 0_{Lq} \end{bmatrix} \geqslant 0. \quad (25)$$

Then the following theorem holds.

Theorem 8. Assume that $H_1(w)$ has full row rank and that $\bar{N}$ has at least one positive eigenvalue. Then the input-output data $u(0), u(1), \ldots, u(T), y(0), y(1), \ldots, y(T)$ are informative for quadratic stabilization if and only if there exist matrices $D \in \mathbb{R}^{Lq\times m}$, $\Phi \in \mathbb{S}^{Lq}$, $\Phi > 0$, and a scalar $\alpha \geqslant 0$ such that the LMI (25) holds. In that case, the feedback controller with coefficient matrix $C := -D^\top \Phi^{-1}$ stabilizes all systems of the form (1) that are compatible with the input-output data. Moreover, the QDF Q_Ψ with $\Psi := \Phi^{-1}$ is a common Lyapunov function for all resulting closed loop systems.

REFERENCES

Berberich, J., Scherer, C., and Allgöwer, F. (2021). Combining prior knowledge and data for robust controller design. *https://arxiv.org/abs/2009.05253v3*.

De Persis, C. and Tesi, P. (2020). Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, 65(3), 909–924.

Kojima, C. and Takaba, K. (2005). A generalized Lyapunov stability theorem for discrete-time systems based on quadratic difference forms. In *Proceedings of the 44th IEEE Conference on Decision and Control*, 2911–2916.

Kojima, C. and Takaba, K. (2006). An LMI condition for asymptotic stability of discrete-time system based on quadratic difference forms. In *Proceedings of the IEEE Conference on Computer Aided Control System Design*, 1139–1143.

Trentelman, H.L., van Waarde, H.J., and Camlibel, M.K. (2020). An informativity approach to the algebraic regulator problem. *https://arxiv.org/abs/2009.01552*.

van Waarde, H.J., Camlibel, M.K., and Mesbahi, M. (2022). From noisy data to feedback controllers: nonconservative design via a matrix S-lemma. *IEEE Transactions on Automatic Control*, 67(1), 162?175.

van Waarde, H.J., Eising, J., Trentelman, H.L., and Camlibel, M.K. (2020). Data informativity: a new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11), 4753–4768.

Willems, J.C. and Trentelman, H.L. (1998). On quadratic differential forms. *SIAM Journal on Control and Optimization*, 36(5), 1703–1749.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On shifted passivity of multi-producer heating networks with storage ⋆⋆

Juan E. Machado [*] **Michele Cucuzzella** [**]
Jacquelien M. A. Scherpen [*]

[*] *Jan C. Willems Center for Systems and Control, ENTEG, Faculty of Science and Engineering, University of Groningen, Nijenborgh 4, 9747 AG Groningen, the Netherlands (e-mail: j.e.machado(j.m.a.scherpen)@rug.nl).*
[**] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy (e-mail: michele.cucuzzella@unipv.it)*

Abstract: We present a nonlinear ODE-based thermo-hydraulic model of a district heating system with multiple heat producers, consumers and storage devices. We analyze the conditions under which the hydraulic and thermal subsystems of the model exhibit shifted passivity properties. For the hydraulic subsystem, our claims on passivity draw on the monotonicity of the vector field associated with the district heating system's flow dynamics, which mainly codifies viscous friction effects on the system's pressures. For the temperature dynamics, we propose a storage function based on the *ectropy function* of a thermodynamic system, recently used in the passivity analysis of heat exchanger networks.

Keywords: Heating networks, modeling, shifted passivity.

1. INTRODUCTION

District heating (DH) has been identified as a key technology to enable the heating sector's potential to reduce greenhouse emissions due to the possibility to seamlessly include environmentally friendly energy sources and storage devices (see, *e.g.*, Lund et al. (2014)). A DH system comprises a network of pipes connecting buildings in a neighborhood, town center or whole city, so that they can be served from varied heat production units (Lund et al., 2014). To further unlock the potential of DH systems, prospective installations will feature multiple, distributed heat sources, *e.g.,* waste-to-energy facilities or solar collectors Lund et al. (2014), promoting as a consequence heat distribution networks of meshed topology, as opposed to the salient tree-like structure of conventional installations with a single heat source (Wang et al., 2017; Vesterlund et al., 2017).

On the one hand, modeling of DH systems with a single heat producer has been addressed, *e.g.,* in De Persis and Kallesoe (2011); Scholten et al. (2015); Hauschild et al. (2020), whereas the multi-producer case has been considered in Wang et al. (2017); Vesterlund et al. (2017); Trip et al. (2019); Alisic et al. (2019). In Wang et al. (2017); Vesterlund et al. (2017), (static) steady-state hydraulic and thermo-hydraulic models are respectively considered for solving operational optimization problems. Dynamic volume storage modeling and control is considered in Trip et al. (2019) for a DH system with multiple storage tanks and neglecting the thermal dynamics. A similar model further considering thermal dynamics is established in Alisic et al. (2019), but it neglects the flow dynamics of the distribution network.

On the other hand, passivity analysis within the context of heating networks has been considered in Mukherjee et al. (2012); Dong et al. (2019). In Mukherjee et al. (2012), a (linear) model to describe the temperature dynamics of a multi-zone building is presented and subsequently shown to be passive via a storage function which is quadratic in the rooms' temperatures. In Dong et al. (2019), a general model of a network of heat exchangers is shown to be *shifted* passive using a novel storage function based on the concept of ectropy (Haddad, 2019). It was mentioned in a previous paragraph that port-Hamiltonian formulations of (single producer) DH system models are presented in Hauschild et al. (2020), thus, passivity follows directly under mild assumptions (van der Schaft and Jeltsema, 2014). [1]

Based on De Persis and Kallesoe (2011); Scholten et al. (2015); Wang et al. (2017); Hauschild et al. (2020) and others (see all the details and proper referencing in Machado

⋆ This research was performed as part of the TOP-UP project (No 91176), which received funding from the Netherlands Organisation for Scientific Research (NWO) and the framework of the joint programming initiative ERA-Net Smart Energy Systems' focus initiative Integrated, Regional Energy Systems, with support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 775970.
⋆⋆This in an extended abstract of the article Machado et al. (2022).

[1] Ectropy is a quadratic function on the total energy of a thermodynamic system and is described in Haddad (2019) as the dual of entropy in the sense that it represents a measure of the tendency of a thermodynamic system to do useful work and grow more organized (see also Willems (2006)). On the other hand, shifted passivity is particularly relevant in these applications by allowing the stability assessment and stabilization of non-trivial equilibria (see Jayawardhana et al. (2007); Monshizadeh et al. (2019b,a)).

et al. (2022)), we present a nonlinear ODE-based model to describe the hydraulic and thermal dynamics of a DH system with multiple heat producers, storage devices and consumers. Moreover, we describe the conditions under which flow and thermal dynamics of the proposed model are shifted passive. Our claims on the passivity of the DH system's flow dynamics are based on the observations made in De Persis and Kallesoe (2011), in the single producer setting, about the monotonicity of the associated vector field. On the other hand, following Dong et al. (2019) (see also Hauschild et al. (2020)), for the thermal dynamics we propose a quadratic storage function based on the total ectropy, extending the results of Dong et al. (2019); Hauschild et al. (2020) to multi-producer systems with storage units.

Notation: The symbol $\mathbb{R}$ denotes the set of real numbers. For a vector $x \in \mathbb{R}^n$, x_i denotes its ith component, *i.e.*, $x = [x_1, \dots, x_n]^\top$; moreover, $\mathbf{sign}(x) = [\mathrm{sign}(x_1), \dots, \mathrm{sign}(x_n)]^\top$, with $\mathrm{sign}(0) = 0$, and $|x| = [|x_1|, \dots, |x_n|]^\top$. An $m \times n$ matrix with all-zero entries is written as $\mathbf{0}_{m\times n}$. An n-vector of ones is written as $\mathbf{1}_n$, whereas the identity matrix of size n is represented by I_n. For any vector $x \in \mathbb{R}^n$, we denote by $\mathrm{diag}(x)$ a diagonal matrix with elements x_i in its main diagonal. For any time-varying signal w, we represent by $\bar{w}$ its steady-state value, if exists. Also, we write time derivatives as $\dot{x}(t)$, and omit the argument t whenever is clear from the context.

2. SYSTEM SETUP

A schematic representation of a DH system with multiple heat producers, consumers and storage tanks is shown in Fig. 1. Producers and consumers are interconnected through a distribution network (DN) with independent supply (hot) and return (cold) layers. The specific composition of producers and consumers is shown in Fig. 2. Note that each producer drains water from the DN's return layer and injects heated water into the supply layer; a converse operation follows for consumers.

A storage tank stores a mixture of hot and cold water perfectly separated by a thermocline, hot layer is on top and the cold one at the bottom. It is assumed that there is no heat or mass exchange between the mixtures. Also, each storage tank has four valves, two at the top and two at the bottom, which are used as inlets and outlets of hot and cold water, respectively. Out of simplicity, we assume that each producer is interfaced to the DN via a storage tank, *i.e.,* each producer drains water from the cold layer of a storage tank and injects it into the storage tank's hot layer. Using the other pair of inlet/outlet valves, each tank drains water from the return layer of the DN and injects it into its cold layer, and at the same time, the storage tank injects the same amount of water from its hot layer into the DN's supply layer.

The overall DH system is viewed as a connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with no self-loops (see, *e.g.*, De Persis and Kallesoe (2011); Wang et al. (2017); Hauschild et al. (2020)). The set of edges $\mathcal{E}$ contains all two-terminal devices (valves, pumps or pipes), and the set of nodes $\mathcal{N}$ contains all junctions as well as the hot and cold layers of each storage tank. The cardinalities of $\mathcal{N}$ and $\mathcal{E}$ are denoted by n_N and n_E, respectively. Taking as reference the sketch in

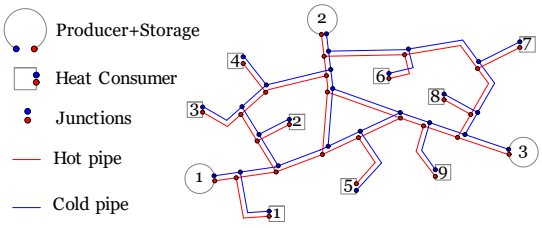


Fig. 1. Sketch based on Wang et al. (2017) of a DH system with 3 heat producers and 9 consumers.

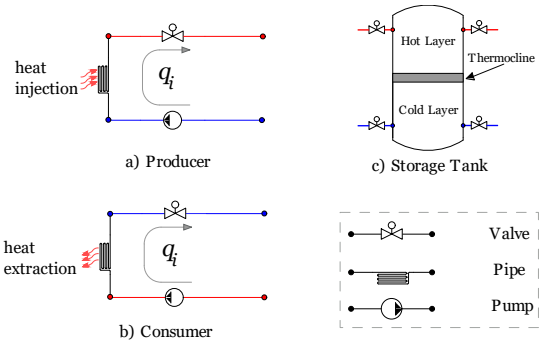


Fig. 2. Topologies of producers, consumers and storage tanks; see Scholten et al. (2015); De Persis and Kallesoe (2011). For producers and consumers, the pipes represent heat exchangers.

Fig. 1, the gray, blue and red lines therein represent edges, whereas colored circles, nodes.

The variables $q_{\mathrm{E},i}$, $V_{\mathrm{E},i}$, $T_{\mathrm{E},i}$ and $p_{\mathrm{E},i}$ denote the flow rate, volume, temperature and pressure of the stream through $i \in \mathcal{E}$. Analogous descriptions follow for the variables $V_{\mathrm{N},k}$, $T_{\mathrm{N},k}$ and $p_{\mathrm{N},k}$, $k \in \mathcal{N}$. Also, we fix an arbitrary orientation to every edge of $\mathcal{G}$. Then, for any $i \in \mathcal{E}$ with end nodes $j, k \in \mathcal{N}$, $j \neq k$, we say that j is the head and k is the tail of i, or viceversa, that j is the tail and k is the head of i. Then, for each node we define the following sets (Hauschild et al., 2020; Krug et al., 2021; Vladimarsson, 2014):

$$\mathfrak{S}_k = \{i \in \mathcal{E} : k \text{ is the tail of } i \in \mathcal{E}\},\ k \in \mathcal{N}, \tag{1a}$$

$$\mathfrak{T}_k = \{i \in \mathcal{E} : k \text{ is the head of } i \in \mathcal{E}\},\ k \in \mathcal{N}. \tag{1b}$$

We define a constant incidence matrix $\mathcal{B}_0$ associated with the arbitrary orientation we have fixed for the DH system's edges, as follows:

$$(\mathcal{B}_0)_{i,j} = \begin{cases} 1, & \text{if node } i \text{ is the head of edge } j, \\ -1, & \text{if node } i \text{ is the tail of edge } j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

For simplicity of exposition, we introduce the preliminary assumption that the orientation of any edge $i \in \mathcal{E}$ matches the direction of the stream through it. That is, if $j, k \in \mathcal{N}$, $j \neq k$, are the tail and head of any $i \in \mathcal{E}$, respectively, then the stream through i is assumed to flow from j to k and we consider that $q_{\mathrm{E},i} \geq 0$.

The following are standing assumptions in this work:

Assumption 1. (i) The density $\rho > 0$ and specific heat $c_{\mathrm{s.h.}} > 0$ of water are spatially uniform and constant in time; for ease of notation we take $\rho = c_{\mathrm{s.h.}} = 1$. (ii) All pipes are cylindrical. (iii) The flow through any edge $i \in \mathcal{E}$ is (spatially) one-dimensional. (iv) Gravitational forces are neglected. (v) The pressure of each $k \in \mathcal{N}$ is spatially uniform and for each tank the pressure of its layers is

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

equal. (vi) Each device (pipe, valve, pump, storage tank, junction) is completely filled with water all the time.

3. HYDRAULIC DYNAMICS

In this section we present a model to describe the dynamic behavior of the hydraulic variables of the DH system.

Under Assumption 1, the equations for mass and momentum balance at each edge $i \in \mathcal{E}$ can be written as follows:

$$q_{\mathrm{E},i} = q^{\mathrm{in}}_{\mathrm{E},i} = q^{\mathrm{out}}_{\mathrm{E},i} \tag{3a}$$

$$p^{\mathrm{in}}_{\mathrm{E},i} - p^{\mathrm{out}}_{\mathrm{E,i}} = J_{\mathrm{E},i}\dot{q}_{\mathrm{E},i} + f_{\mathrm{E},i}(q_{\mathrm{E},i}) - w_{\mathrm{E},i}, \tag{3b}$$

$$f_{\mathrm{E},i}(q_{\mathrm{E}}, i) = \theta_{\mathrm{E},i}|q_{\mathrm{E},i}|q_{\mathrm{E},i}, \tag{3c}$$

where $q^{\mathrm{in}}_{\mathrm{E},i}$, $q^{\mathrm{out}}_{\mathrm{E},i}$ and $p^{\mathrm{in}}_{\mathrm{E},i}$, $p^{\mathrm{out}}_{\mathrm{E},i}$ are the pipe's inlet-outlet flow and pressure pairs. Note that $V_{\mathrm{E},i}$ is constant all the time. If $i \in \mathcal{E}$ is a pipe, then $J_{\mathrm{E},i} = (\rho\ell_{\mathrm{E},i})/A_{\mathrm{E},i} > 0$, where $\ell_{\mathrm{E},i}$ and $A_{\mathrm{E},i}$ are the pipe's length and cross-section area; also $\theta_{\mathrm{E},i} > 0$ depends on the pipe's friction factor and diameter. If i is a valve, then $J_{\mathrm{E},i} = V_{\mathrm{E},i} = w_{\mathrm{E},i} = 0$ and $\theta_{\mathrm{E},i} > 0$. The latter parameter is constant and represents the valve's friction coefficient. If i is a pump, then $J_{\mathrm{E},i} = V_{\mathrm{E},i} = 0$ and $w_{\mathrm{E},i}$ is the pressure difference produced between its terminals.

On the other hand, for each node $k \in \mathcal{N}$ the following constraints are considered:

$$\dot{V}_{\mathrm{N},k} = \sum_{i\in\mathfrak{T}_k} q_{\mathrm{E},i} - \sum_{i\in\mathfrak{S}_k} q_{\mathrm{E},i}, \tag{4a}$$

$$p_{\mathrm{N},k} = p^{\mathrm{in}}_{\mathrm{E},i},\ i \in \mathfrak{S}_k, \quad p_{\mathrm{N},k} = p^{\mathrm{out}}_{\mathrm{E},i},\ i \in \mathfrak{T}_k, \tag{4b}$$

where the sets $\mathfrak{S}_k$ and $\mathfrak{T}_k$ are defined in (1). Equation (4a) represents mass balance and (4b) guarantees pressure consistency at each node. Note that under Assumption 1.(vi), the right hand side of (4a) must be zero if $k \in \mathcal{N}$ is a simple junction. Moreover, if $a, b \in \mathcal{N}$ represent the hot and cold layer of a given storage tank, then $p_{\mathrm{N},a} = p_{\mathrm{N},b}$ and $V_{\mathrm{N},a} + V_{\mathrm{N},b} = V^{\max}$, for some $V^{\max} > 0$ denoting the storage tank's capacity. These additional constraints complement (4).

We note that (3) and (4) can be represented in vector form as the following DAE:

$$-B_0^\top p_{\mathrm{n}} = \mathrm{diag}(J_{\mathrm{E}})\dot{q}_{\mathrm{E}} + f_{\mathrm{E}}(q_{\mathrm{E}}) - w_{\mathrm{E}}, \tag{5a}$$

$$\dot{V}_{\mathrm{n}} = 0 = B_0 q_{\mathrm{E}}, \tag{5b}$$

where p_{N}, V_{N} are the pressure and volume vectors of the reduced DH system's graph $G = (N, \mathcal{E})$, resulting from $\mathcal{G}$ when for each storage tank we merge its hot and cold layers. Also, B_0 is an incidence matrix for G.

Following De Persis and Kallesoe (2011), a set of independent flows can be identified from which the entire hydraulic state of the DH system can be determined. These flows are associated with a selected collection of *pipes* that generate fundamental loops of G.

Theorem 1. There exists a collection $\mathcal{C} \subset \mathcal{E}$ of n_{f} pipes whose flows $q_{\mathrm{f},i}$ are independent variables. All system flows can be computed as $q_{\mathrm{E}} = F^\top q_{\mathrm{f}}$, where F is the (full rank) fundamental loop matrix associated with $\mathcal{C}$ (and G). Moreover, the following claims hold true:

(I) If there is an independently controlled pump with pressure $w_{\mathrm{f},i}$ adjacent to each pipe in $\mathcal{C}$, then q_{f} is governed by the dynamics

$$\mathcal{J}_{\mathrm{f}}\dot{q}_{\mathrm{f}} = -f_{\mathrm{f}}(q_{\mathrm{f}}) + w_{\mathrm{f}} + B_{\mathrm{b}}w_{\mathrm{b}}, \tag{6}$$

where $\mathcal{J}_f = F\mathrm{diag}(J_{\mathrm{E}})F^\top > 0$ and $f_{\mathrm{f}}(q_{\mathrm{f}}) = Ff_{\mathrm{E}}(F^\top q_{\mathrm{f}})$ is a monotone function. Also, $B_{\mathrm{b}}w_{\mathrm{b}}$, with $(B_{\mathrm{b}})_{\alpha,\beta} \in \{-1, 0, 1\}$, codifies the effect on q_{f} of any other pump in the system (here we assume they provide a constant pressure difference).

(II) There exists $W \in \mathbb{R}^{n_{\mathrm{ST}}\times n_{\mathrm{f}}}$, with entries in $\{0, 1\}$ and n_{ST} being the number of storage tanks, such that the dynamics of the storage tanks' hot layer volume are given by

$$\dot{V}_{\mathrm{sh}} = Wq_{\mathrm{f}}. \tag{7}$$

Also, $\dot{V}_{\mathrm{sc}} = -Wq_{\mathrm{f}}$ holds for the cold layers' volumes.

(III) If w_{b} is constant, then (6) is shifted passive with passive output q_{f} and storage function $\mathcal{S}_{\mathrm{f}}(q_{\mathrm{f}}) = \frac{1}{2}(q_{\mathrm{f}} - \bar{q}_{\mathrm{f}})^\top\mathcal{J}_{\mathrm{f}}(q_{\mathrm{f}} - \bar{q}_{\mathrm{f}})$. It follows that $\dot{\mathcal{S}}_{\mathrm{f}} \le (w_{\mathrm{f}} - \bar{u}_{\mathrm{f}})^\top(q_{\mathrm{f}} - \bar{q}_{\mathrm{f}})$ holds for all time and for any equilibrium pair $(\bar{w}_{\mathrm{f}}, \bar{q}_{\mathrm{f}})$ of (6).

4. TEMPERATURE DYNAMICS

We present now a model to describe the dynamic behavior of the DH system's temperatures. More details, as well as the standing assumptions behind it are described in Machado et al. (2022). For ease of notation, we assume that the orientation of each $i \in \mathcal{E}$ matches the direction of the stream through it, *i.e.*, $q_{\mathrm{E},i} \ge 0$ all the time.

Let $i \in \mathcal{E}$ either be a pipe, a valve of a pump. Then, the energy balance at i is equivalent to:

$$\begin{aligned} V_{\mathrm{E},i}\dot{T}_{\mathrm{E},i} = q_{\mathrm{E},i}\left(T^{\mathrm{in}}_{\mathrm{E},i} - T^{\mathrm{out}}_{\mathrm{E},i}\right) + \alpha_{\mathrm{pr},i}P_{\mathrm{pr},i} - \beta_{\mathrm{c},i}P_{\mathrm{c},i} \\ + q_{\mathrm{E},i}f_{\mathrm{E},i}(q_{\mathrm{E}}, i), \end{aligned} \tag{8}$$

where $T^{\mathrm{in}}_{\mathrm{E},i}$ ($T^{\mathrm{out}}_{\mathrm{E},i}$) is the temperature at the inlet (outlet) of the pipe. If i is associated with the heat exchanger of a producer (consumer), then $P_{\mathrm{pr},i}$ ($P_{\mathrm{c},i}$) is the heat injection (extraction) into (from) the DH system by the producer (consumer). Also, each $P_{\mathrm{pr},i}$ is a control input, whereas each $P_{\mathrm{c},i}$ is a disturbance. The term $q_{\mathrm{E},i}f_{\mathrm{E},i}(q_{\mathrm{E},i})$ represents heat dissipation due to frictional forces.

Now let $k \in \mathcal{N}$ be an arbitrary node of the DH system. Then, the energy balance at k can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(V_{\mathrm{N},k}T_{\mathrm{N},k}\right) = \sum_{j\in\mathfrak{T}_k} q_{\mathrm{E},j}T^{\mathrm{out}}_{\mathrm{E},j} - \sum_{j\in\mathfrak{S}_k} q_{\mathrm{E},j}T^{\mathrm{in}}_{\mathrm{E},j}. \tag{9}$$

The term in the left-hand side represents the rate of change of the thermal energy stored at node k whereas the terms in the right-hand side are the sum of the thermal energies of the streams that target and source from k, respectively.

Based on the (upwind) semi-discretization scheme discussed in Hauschild et al. (2020) and on the nodal constraints described in Krug et al. (2021), we complement (8) and (9) with the following constraints for each $i \in \mathcal{E}$ and $j \in \mathcal{N}$:

$$T^{\mathrm{in}}_{\mathrm{E},i} = T_{\mathrm{N},j},\ \forall i \in \mathfrak{S}_j \subset \mathcal{E}, \text{ and } \quad T^{\mathrm{out}}_{\mathrm{E},i} = T_{\mathrm{E},i}. \tag{10a}$$

It follows that (9) can be written in a the following, equivalent form:

$$V_{\mathrm{N},k}\dot{T}_{\mathrm{N},k} = \sum_{j\in\mathfrak{T}_k} q_{\mathrm{E},j}T_{\mathrm{E},j} - \left(\sum_{j\in\mathfrak{T}_k} q_{\mathrm{E},j}\right)T_{\mathrm{N},k},$$

where we have used (4a). Note that $V_{\mathrm{N},k}$ is constant for most of the nodes, with the exception of the layers of storage tanks.

By defining $\mathcal{T} = \frac{1}{2}(\mathcal{B}_0 + |\mathcal{B}_0|)$ and $\mathcal{S} = \frac{1}{2}(\mathcal{B}_0 - |\mathcal{B}_0|)$, then the system (8), (10) and (11) can be written as follows:

$$\mathrm{diag}(V_{\mathrm{E}}, V_{\mathrm{N}}) \begin{bmatrix} \dot{T}_{\mathrm{E}} \\ \dot{T}_{\mathrm{N}} \end{bmatrix} = \mathcal{A}(q_{\mathrm{E}}) \begin{bmatrix} T_{\mathrm{E}} \\ T_{\mathrm{N}} \end{bmatrix} + B_{\mathrm{pr}} P_{\mathrm{pr}} - B_{\mathrm{c}} P_{\mathrm{c}} + \begin{bmatrix} \mathrm{diag}(q_{\mathrm{E}}) f_{\mathrm{E}}(q_{\mathrm{E}}) \\ 0_{n_{\mathrm{N}} \times n_{\mathrm{E}}} \end{bmatrix}, \tag{11}$$

where

$$\mathcal{A}(q_{\mathrm{E}}) = \begin{bmatrix} -\mathrm{diag}(q_{\mathrm{E}}) & \mathrm{diag}(q_{\mathrm{E}})\mathcal{S}^\top \\ \mathcal{T}\mathrm{diag}(q_{\mathrm{E}}) & -\mathrm{diag}(\mathcal{T} q_{\mathrm{E}}) \end{bmatrix}. \tag{12}$$

Also, P_{pr} (P_{c}) collects the producers (consumers) heat injections (extractions) to (from) the DN, then B_{pr} (B_{c}) is a suitable constant matrix with entries in $\{0, 1\}$.

Consider the following:

Theorem 2. Consider the system (6) and (12) and assume that P_{c} is constant. Then, the following claims hold true:

(I) The system is cyclo-dissipative (van der Schaft, 2021) with the total energy $\mathcal{H} = \frac{1}{2} q_{\mathrm{f}}^\top \mathcal{J}_{\mathrm{f}} q_{\mathrm{f}} + \sum_{i \in \mathcal{E}} V_{\mathrm{E},i} T_{\mathrm{E},i} + \sum_{i \in \mathcal{N}} V_{\mathrm{N},i} T_{\mathrm{N},i}$ as storage function and $w_{\mathrm{f}}^\top q_{\mathrm{f}} + \mathbf{1}^\top P_{\mathrm{pr}} - \mathbf{1}^\top P_{\mathrm{c}}$ as supply rate.

(II) Assume that q_{E} is at equilibrium. Then $\mathcal{A}(q_{\mathrm{E}})$ is a Kirchhoff Convection Matrix, which implies that $\mathcal{A}(q_{\mathrm{E}}) \leq 0$. It follows that (11) is shifted passive with passive output $T_{\mathrm{pr}} := B_{\mathrm{pr}}^\top T_{\mathrm{E}}$ and storage function $\mathcal{S}_{\mathrm{th}}(T_{\mathrm{E}}) = \frac{1}{2}(T_{\mathrm{E}} - \bar{T}_{\mathrm{E}})^\top \mathrm{diag}(V_{\mathrm{E}})(T_{\mathrm{E}} - \bar{T}_{\mathrm{E}}) + \frac{1}{2}(T_{\mathrm{N}} - \bar{T}_{\mathrm{N}})^\top \mathrm{diag}(V_{\mathrm{N}})(T_{\mathrm{N}} - \bar{T}_{\mathrm{N}})$. That is, $\dot{\mathcal{S}}_{\mathrm{th}} \leq (P_{\mathrm{pr}} - \bar{P}_{\mathrm{pr}})^\top (T_{\mathrm{pr}} - \bar{T}_{\mathrm{pr}})$ holds for all time and for any equilibrium pair $(\bar{P}_{\mathrm{pr}}, \bar{T}_{\mathrm{E}}, \bar{T}_{\mathrm{N}})$ of (11).

We note that the storage function $\mathcal{S}_{\mathrm{th}}$ in Theorem 2 is based on the shifted ectropy (Haddad, 2019), which is quadratic in the total energy of the system (Haddad, 2019, Chapter 3). Also, the passive output T_{pr} stacks the temperatures of the heated water streams that each producer injects into the DH system, respectively.

5. CONCLUSION

Invoking conservation laws and graph theoretic tools we have derived a thermo-hydraulic model of a multi-producer DH system and established that it is cyclo-dissipative. Also, its hydraulic and thermal subsystems are shifted passive under certain conditions.

In the talk we will discuss and present numerical simulations about how the established shifted passivity properties can be used in the synthesis of *decentralized* controllers with closed-loop stability guarantees. We will also discuss current modeling extensions aimed at designing optimal control strategies for the real time minimization of meaningful cost functions related with consumers' thermal comfort.

REFERENCES

Alisic, R., Paré, P., and Sandberg, H. (2019). Modeling and Stability of Prosumer Heat Networks. *IFAC-PapersOnLine*, 52(20), 235–240. doi:10.1016/j.ifacol.2019.12.164.

De Persis, C. and Kallesoe, C. (2011). Pressure regulation in nonlinear hydraulic networks by positive and quantized controls. *IEEE Transactions on Control Systems Technology*, 19(6), 1371–1383. doi:10.1109/TCST.2010.2094619.

Dong, Z., Li, B., and Huang, X. (2019). Passivity-based control of heat exchanger networks. 6531–6536.

Haddad, W.M. (2019). *A Dynamical Systems Theory of Thermodynamics*. Princeton University Press.

Hauschild, S., Marheineke, N., Mehrmann, V., Mohring, J., Badlyan, A., Rein, M., and Schmidt, M. (2020). Port-hamiltonian modeling of district heating networks. In *Progress in Differential-Algebraic Equations II*, 333–355. Springer International Publishing, Cham.

Jayawardhana, B., Ortega, R., García-Canseco, E., and Castaños, F. (2007). Passivity of nonlinear incremental systems: Application to PI stabilization of nonlinear RLC circuits. *Systems and Control Letters*, 56, 618–622. doi:10.1016/j.sysconle.2007.03.011.

Krug, R., Mehrmann, V., and Schmidt, M. (2021). Nonlinear optimization of district heating networks. *Optimization and Engineering*, 22(2), 783–819.

Lund, H., Werner, S., Wiltshire, R., Svendsen, S., Thorsen, J.E., Hvelplund, F., and Mathiesen, B.V. (2014). 4th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems. *Energy*, 68, 1–11. doi:10.1016/j.energy.2014.02.089. URL `http://dx.doi.org/10.1016/j.energy.2014.02.089`.

Machado, J.E., Cucuzzella, M., and Scherpen, J.M. (2022). Modeling and passivity properties of multi-producer district heating systems. *Automatica*, 142.

Monshizadeh, N., Monshizadeh, P., Ortega, R., and van der Schaft, A. (2019a). Conditions on shifted passivity of port-Hamiltonian systems. *Systems and Control Letters*, 123, 55–61. URL `https://doi.org/10.1016/j.sysconle.2018.10.010`.

Monshizadeh, P., Machado, J., Ortega, R., and van der Schaft, A. (2019b). Power-controlled Hamiltonian systems: Application to electrical systems with constant power loads. *Automatica*, 109.

Mukherjee, S., Mishra, S., and Wen, J.T. (2012). Building temperature control: A passivity-based approach. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 6902–6907. IEEE.

Scholten, T., De Persis, C., and Tesi, P. (2015). Modeling and control of heat networks with storage: The single-producer multiple-consumer case. *IEEE Transactions on Control Systems Technology*, 25(2), 2242–2247. doi:10.1109/ECC.2015.7330872.

Trip, S., Scholten, T., and De Persis, C. (2019). Optimal regulation of flow networks with transient constraints. *Automatica*, 104, 141–153. doi:10.1016/j.automatica.2019.02.046.

van der Schaft, A. (2021). Classical thermodynamics revisited: A systems and control perspective. *IEEE Control Systems Magazine*, 41(5), 32–60.

van der Schaft, A. and Jeltsema, D. (2014). *Port-Hamiltonian systems theory: An introductory overview*, volume 1. Foundations and Trends in Systems and Control.

Vesterlund, M., Toffolo, A., and Dahl, J. (2017). Optimization of multi-source complex district heating network, a case study. *Energy*, 126, 53–63.

Vladimarsson, P. (2014). District heat distribution networks. In *Short Course VI on Utilization of Low- and Medium-Enthalpy Geothermal Resources and Financial Aspects of Utilization*. UNU-GTP and LaGeo.

Wang, Y., You, S., Zhang, H., Zheng, W., Zheng, X., and Miao, Q. (2017). Hydraulic performance optimization of meshed district heating network with multiple heat sources. *Energy*, 126, 603–621. doi:10.1016/j.energy.2017.03.044. URL `http://dx.doi.org/10.1016/j.energy.2017.03.044`.

Willems, J.C. (2006). Thermodynamics: A Dynamical Systems Approach—W. M. Haddad, V. S. Chellaboina, and S. Nersesnov. *IEEE Transactions on Automatic Control*, 51, 1217–1225. Book review.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Extended Kalman filter observer design for semilinear infinite-dimensional systems ⋆

Sepideh Afshar * **Fabian Germ** ** **Kirsten Morris** ***

* *Dept. of Radiology, Harvard Medical School, Massachusetts General Hospital, Boston, MA, USA. (email: safshar1@mgh.harvard.edu)*
** *School of Mathematics, University of Edinburgh, Edinburgh, UK (email: f.germ@sms.ed.ac.uk)*
*** *Department of Applied Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada (email: kmorris@uwaterloo.ca)*

Abstract:
A semilinear infinite-dimensional system with a disturbance input is considered. The observation is modelled by an affine linear map with a different disturbance. An observer, based on the extended Kalman filter (EKF), is constructed and its well-posedness is proven under mild conditions. Moreover, local exponential stability of the error dynamics is shown. Thus, if the error in the initial condition is small enough, the estimation error converges to zero. This is a first generalization of the EKF to infinite-dimensional systems. Since only detectability, not observability, is assumed, this result is new even for finite-dimensional systems. An implementation is provided for a magnetic drug-delivery system and numerical results support the effectiveness of the observer.

Keywords: estimation, extended Kalman filter (EKF), infinite-dimensional system, semilinear partial differential equations

AMS Classification: 93B99 , 93C20

1. INTRODUCTION

In many physical applications, the system's state varies with spatial variables as well as time. The state of such systems is modelled by partial differential equations and evolves on an infinite-dimensional space, and so they are an important class of infinite-dimensional systems, as are systems modelled by delay-differential equations. The full state of these systems cannot be measured. As for finite-dimensional systems, a system, referred to as an observer or estimator, can be designed to estimate the state using the mathematical model and the measurements provided by sensors.

For linear systems, the Kalman filter (KF) minimizes the variance of the error under certain assumptions on the disturbances. The observer can be calculated through solution of a Riccati equation. The Kalman filter is widely used and was extended to infinite-dimensional linear systems in the 1970's; see the review papers Curtain (1975) and **?**. This theory was recently extended to time-varying infinite-dimensional systems, see Wu et al. (2015).

Due to its success in a wide range of applications, an extension of the KF to nonlinear systems, the extended Kalman filter (EKF), was developed for finite-dimensional systems. The EKF design is based on a linear approximation of the system around the estimated state. The linearized system is used to derive a Riccati equation and this is used to calculate the observer gain; e.g. Simon (2006); Grewal and Andrews (2011). This method is widely used; see for example, Reif and Unbehauen (1999); Reif et al. (1999, 2000); Kai et al. (2010, 2011); Einicke and White (1999). However, although this method may work well, it is well known that it may lead to divergent error estimates. Convergence of the estimation error for EKF depends on the size of the nonlinearity and the initial condition, see for instance, Liang (1983); Ribeiro (2004).

Further convergence results include local exponential convergence of the estimation error in Baras et al. (1988) or Reif et al. (1998) under certain conditions, as well as in Ahrens and Khalil (2007) using the normal form of the governing ordinary differential equations.

Observers for nonlinear infinite-dimensional systems are often designed using a finite-dimensional approximation of the system. This enables the use of techniques for nonlinear finite-dimensional systems. For an example thereof, using the EKF, see Rigatos et al. (2017) and Afshar et al. (2018).

There are some studies for nonlinear infinite-dimensional systems where the observer is designed directly using the infinite-dimensional system equations. To name some select examples, see the second-order sliding mode observer in Miranda et al. (2012), an observer with correction by a linear output error injection in Bitzer and Zeitz (2002), or spatially-distributed linear output injection in Efe* et al. (2005) for a one-dimensional nonlinear Burgers' equation.

⋆ Financial support of Natural Sciences and Engineering Research Council of Canada (NSERC) and of the U.S. AFOSR under Grant FA9550-16-1-0061 for this research is gratefully acknowledged.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The EKF is formally shown here to be well-posed for a class of semilinear infinite-dimensional systems with bounded observation. As for a finite-dimensional EKF, the observer dynamics are a copy of the original system's dynamics with an injection gain defined by the solution of an Riccati equation. Since the Riccati equation is coupled with the observer equation, conventional results in the literature including Curtain and Pritchard (1976) for existence of solutions to the Riccati equation cannot be directly used. This is due to the fact that for linear equations the Riccati equation does not depend on the state of the system. In our, nonlinear, case, such a dependence still remains after linearizing the system, making the analysis more involved.

Also, for sufficiently small initial error, and smooth nonlinearity, the error dynamics are exponentially stable. Similar results for finite-dimensional systems; see Elizabeth and Jothilakshmi (2015); Alonge et al. (2014); Reif et al. (1999, 2000) assumed uniform observability . Here only detectability is assumed so the results are new for finite-dimensional systems. The estimation error is bounded in presence of disturbances.

For implementation, the infinite-dimensional EKF must be approximated using some method. The paper concludes with illustration of implementation using a finite-element method for estimation of concentration in a magnetic drug delivery system.

The proof of well-posedness was done in the thesis Germ (2019) for nonlinearities without time dependence, and briefly sketched in the conference paper Afshar et al. (2020). Complete details of the results described in this talk are in **?**.

2. PROBLEM STATEMENT

Let $\mathcal{H}$ be a Hilbert space and let $\boldsymbol{A} : \mathcal{D}(\boldsymbol{A}) \to \mathcal{H}$ be a linear operator that generates a C_0-semigroup $\boldsymbol{T}(t)$ on $\mathcal{H}$ and $\boldsymbol{F} : \mathcal{H} \times [0, t_f] \to \mathcal{H}$ be strongly continuous in time, satisfying $\boldsymbol{F}(0, t) = 0$ for every t.

We consider the semilinear evolution system

$$\begin{aligned} \frac{\partial \boldsymbol{z}(t)}{\partial t} &= \boldsymbol{A}\boldsymbol{z}(t) + \boldsymbol{F}(\boldsymbol{z}(t), t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{G}\boldsymbol{\omega}(t), \\ \boldsymbol{z}(0) &= \boldsymbol{z}_0 \in \mathcal{H}, \end{aligned} \tag{1}$$

where for Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, $\boldsymbol{u}(t) \in \mathcal{C}([0, t_f], \mathcal{H}_1)$ is the control input, $\boldsymbol{\omega}(t) \in \mathcal{C}([0, t_f], \mathcal{H}_2)$ is the input disturbance and $\boldsymbol{B} \in \mathcal{L}(\mathcal{H}_1, \mathcal{H})$, $\boldsymbol{G} \in \mathcal{L}(\mathcal{H}_2, \mathcal{H})$, with $\mathcal{L}(\mathcal{A}, \mathcal{B})$ denoting the space of linear and bounded operators from $\mathcal{A}$ to $\mathcal{B}$. We refer to $\boldsymbol{z}(t)$ as the state of the system (1).

Assumption 1. The operator $\boldsymbol{F}$ admits a Fréchet-derivative $\boldsymbol{DF}(\cdot, \cdot)$ such that for a constant $\delta_{DF} > 0$ we have $\|\boldsymbol{DF}(x, t)\| \leq \delta_{DF}$ for all $(x, t) \in \mathcal{H} \times [0, t_f]$, and for every $\delta > 0$ there exists a Lipschitz constant $\iota_{DF} > 0$ such that for all $\|\boldsymbol{x} - \boldsymbol{y}\| < \delta$ and all $t \in [0, t_f]$,

$$\|\boldsymbol{DF}(\boldsymbol{x}, t) - \boldsymbol{DF}(\boldsymbol{y}, t)\| \leq \iota_{DF}\|\boldsymbol{x} - \boldsymbol{y}\|.$$

The disturbance $\boldsymbol{\omega}$ and control $\boldsymbol{u}$ may be lumped as

$$\boldsymbol{B}_d = [\boldsymbol{B}, \boldsymbol{G}], \quad \boldsymbol{u}_d^T(t) = [\boldsymbol{u}^T(t), \boldsymbol{\omega}^T(t)].$$

The state-equation for $\boldsymbol{z}$ in system (1) then becomes

$$\begin{aligned} \frac{\partial \boldsymbol{z}(t)}{\partial t} &= \boldsymbol{A}\boldsymbol{z}(t) + \boldsymbol{F}(\boldsymbol{z}(t), t) + \boldsymbol{B}_d\boldsymbol{u}_d(t), \\ z(0) &= z_0 \in \mathcal{H}. \end{aligned} \tag{2}$$

With disturbance $\boldsymbol{\eta}(t) \in \mathcal{C}([0, t_f], \mathbb{R}^p)$, $p \geq 1$, let

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{z}(t) + \boldsymbol{\eta}(t)$$

be the system measurement, where $\boldsymbol{C} \in \mathcal{L}(\mathcal{H}, \mathbb{R}^p)$.

Our objective is to design an observer for the system (2). Most generally, an observer is a dynamical system with state $\hat{\boldsymbol{z}}(t)$ such that, in the absence of disturbances,

$$\lim_{t\to\infty} \|\boldsymbol{z}(t) - \hat{\boldsymbol{z}}(t)\| = 0.$$

In our case, as is common, the observer dynamics contain a copy of the system's dynamics and a feedback term that corrects for the error between the predicted observation, $\boldsymbol{C}\hat{\boldsymbol{z}}$, and the actual observation, $\boldsymbol{y}$. The general form of the observer is

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{z}}(t)}{\partial t} &= \boldsymbol{A}\hat{\boldsymbol{z}}(t) + \boldsymbol{F}(\hat{\boldsymbol{z}}(t), t) + \boldsymbol{B}\boldsymbol{u}(t) + L(t)[\boldsymbol{y}(t) - \boldsymbol{C}\hat{\boldsymbol{z}}(t)] \\ \hat{\boldsymbol{z}}(0) &= \hat{\boldsymbol{z}}_0 \in \mathcal{H}, \end{aligned} \tag{3}$$

where $L(t)$, referred to as observer gain, needs to be selected so that in the absence of disturbances $\boldsymbol{\omega}(t)$ and $\boldsymbol{\eta}(t)$, $\hat{\boldsymbol{z}}(t) \to \boldsymbol{z}(t)$.

3. OBSERVER DESIGN

The problem of observer design for linear systems has been well studied. The most widely known and used approach is the Kalman filter. Consider, instead of (2), the linear system

$$\begin{aligned} \frac{\partial \boldsymbol{z}(t)}{\partial t} &= \tilde{\boldsymbol{A}}(t)\boldsymbol{z}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{\omega}(t). \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{z}(t) + \boldsymbol{\eta}(t). \end{aligned} \tag{4}$$

where $\tilde{\boldsymbol{A}}(t)$ generates an evolution operator $\boldsymbol{U}(t, s)$,and $\boldsymbol{\omega}(t)$ and $\boldsymbol{\eta}(t)$ are process and output disturbance respectively.

Assumption 2. Let linear operators $\boldsymbol{P}_0 \in \mathcal{L}(\mathcal{H})$, $\boldsymbol{W}(t) \in C([0, t_f], \mathcal{L}(\mathcal{H}))$ and $\boldsymbol{R}(t) \in C([0, t_f], \mathcal{L}(\mathbb{R}^p))$ be self-adjoint. Let $\boldsymbol{P}_0$ be positive definite and $\boldsymbol{W}(t)$ be non-negative definite for all $t \in [0, t_f]$. The operator $\boldsymbol{R}(t)$ is uniformly coercive; meaning that there exists a $\delta_0 > 0$ such that for every $\boldsymbol{w} \in \mathbb{R}^p$ and $t \in [0, t_f]$, $(\boldsymbol{w}, \boldsymbol{R}(t)\boldsymbol{w}) \geq \delta_0\|\boldsymbol{w}\|^2$.

Note that since $\boldsymbol{R}(t)$ is self-adjoint, coercive and bounded for each t, it follows that for all $t \in [0, t_f]$, it has a self-adjoint bounded inverse $\boldsymbol{R}^{-1}(t) \in \mathcal{C}([0, t_f], \mathcal{H})$.

Linear integral Riccati equations are defined in the following theorem. For the proof we refer to (Curtain and Pritchard, 1976, Theorem 3.1 & 3.3) or (Curtain, 1976, Theorem 2.1).

Theorem 3. Let $\boldsymbol{U}(t, s)$ be an evolution operator on $\Delta(t_f) := \{(s, t) \in [0, t_f]^2, 0 \leq s \leq t \leq t_f\}$. Under Assumption 2, the following integral Riccati equation

$$\begin{aligned} \boldsymbol{P}(t)\boldsymbol{w} =& \boldsymbol{U}_P(t, 0)\boldsymbol{P}_0\boldsymbol{U}^*(t, 0)\boldsymbol{w} \\ &+ \int_0^t \boldsymbol{U}_P(t, s)\boldsymbol{W}(s)\boldsymbol{U}^*(t, s)\boldsymbol{w}ds, \quad \boldsymbol{w} \in \mathcal{H}, \end{aligned} \tag{5}$$

where the perturbed evolution operator

$$\boldsymbol{U}_P(t,s)\boldsymbol{w} = \boldsymbol{U}(t,s)\boldsymbol{w} - \int_s^t \boldsymbol{U}(t,r)\boldsymbol{P}(r)\boldsymbol{C}^*\boldsymbol{R}^{-1}(r)\boldsymbol{C}\boldsymbol{U}_P(r,s)\boldsymbol{w}dr, \tag{6}$$

$\boldsymbol{w} \in \mathcal{H}$, admits a unique, positive definite, self-adjoint solution $\boldsymbol{P}(t) \in \mathcal{C}([0,t_f],\mathcal{L}(\mathcal{H}))$.

This solution $\boldsymbol{P}(t)$ defines an observer gain,

$$L_P(t) = \boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{R}^{-1}(t).$$

This defines observer dynamics for (4) are

$$\frac{\partial \hat{\boldsymbol{z}}_P(t)}{\partial t} = \tilde{\boldsymbol{A}}(t)\hat{\boldsymbol{z}}_P(t) + \boldsymbol{B}\boldsymbol{u}(t) + L_P(t)(\boldsymbol{y}(t) - C\boldsymbol{z}(t)). \tag{7}$$

In the case that $\boldsymbol{\omega}(t)$ and $\boldsymbol{\eta}(t)$ are process and sensor noises with covariances $\boldsymbol{W}(t)$ and $\boldsymbol{R}(t)$ respectively and the covariance of the initial condition $\hat{\boldsymbol{z}}(0)$ is $\boldsymbol{P}_0$ then the observer gain $L_P(t)$ and corresponding estimate $\hat{\boldsymbol{z}}(t)$ are optimal in a sense that $\hat{\boldsymbol{z}}(t)$ minimizes the error covariance Curtain (1976). This observer is the Kalman filter. For details, see Curtain and Pritchard (1978) and for recent work on time-varying systems, Wu et al. (2015).

The linearization of (2) will be used to define a integral Riccati equation similar to (5). The solution $\boldsymbol{P}$ defines the observer gain. In the case of finite-dimensional systems, this approach is known as an extended Kalman filter (EKF) and this terminology will be used here.

First, the linearization of the system is defined. For this purpose, for $\hat{\boldsymbol{z}}(t) \in C([0,t_f],\mathcal{H})$, at time t the Fréchet-derivative of $\boldsymbol{F}(\cdot,t)$, denoted by $\boldsymbol{DF}(\cdot,t) : \mathcal{H} \to \mathcal{L}(\mathcal{H})$,

$$\boldsymbol{DF}(\hat{\boldsymbol{z}}(t),t) = \frac{\partial \boldsymbol{F}(z,t)}{\partial \boldsymbol{z}} \mid_{\boldsymbol{z}=\hat{\boldsymbol{z}}(t)} . \tag{8}$$

Linearizing the system (2) around $\hat{\boldsymbol{z}}(t)$ yields

$$\frac{\partial \boldsymbol{z}(t)}{\partial t} = \boldsymbol{A}\boldsymbol{z}(t) + \boldsymbol{F}(\hat{\boldsymbol{z}}(t),t) + \boldsymbol{DF}(\hat{\boldsymbol{z}}(t),t)[\boldsymbol{z}(t) - \hat{\boldsymbol{z}}(t)] + \boldsymbol{B}_d\boldsymbol{u}_d(t). \tag{9}$$

To obtain EKF equations, the solution to the integral Riccati equations for the linear system (9) is needed. Since $\boldsymbol{U}$ in (6) is now given by, for $\boldsymbol{w} \in \mathcal{H}$,

$$\boldsymbol{U}(t,s)\boldsymbol{w} = \boldsymbol{T}(t-s)\boldsymbol{w} - \int_s^t \boldsymbol{T}(t-r)\boldsymbol{DF}(\hat{\boldsymbol{z}}_P(r),r)\boldsymbol{U}(r,s)\boldsymbol{w}dr, \tag{10}$$

these will contain the Fréchet-derivative $\boldsymbol{DF}(\hat{\boldsymbol{z}}_P(t),t)$, a possibly nonlinear function of the observer state $\hat{\boldsymbol{z}}_P(t)$. Therefore recent results on the Riccati equation do not provide well-posedness for the coupled system.

For a bounded linear operator $\boldsymbol{P}(t) \in C([0,t_f];\mathcal{L}(\mathcal{H}))$ and $\hat{\boldsymbol{z}}_0 \in \mathcal{H}$, the mild solution $\hat{\boldsymbol{z}}_P(t)$ to (7) is

$$\hat{\boldsymbol{z}}_P(t) = \boldsymbol{T}(t)\hat{\boldsymbol{z}}_0 + \int_0^t \boldsymbol{T}(t-s)\big(\boldsymbol{F}(\hat{\boldsymbol{z}}_P(s),s) + \boldsymbol{B}\boldsymbol{u}(s)\big)ds + \int_0^t \boldsymbol{T}(t-s)\boldsymbol{P}(s)\boldsymbol{C}^*\boldsymbol{R}^{-1}(s)[\boldsymbol{y}(s) - \boldsymbol{C}\hat{\boldsymbol{z}}_P(s)]ds. \tag{11}$$

The following is the first of our main results.

Theorem 4. Let Assumptions 1 and 2 hold. For any $\boldsymbol{u}(t) \in C([0,t_f],\mathcal{H}_1)$, $y(t) \in C([0,t_f],\mathbb{R}^p)$ and $\hat{\boldsymbol{z}}_0 \in \mathcal{H}$ there exist $\hat{\boldsymbol{z}}_P(t) \in C([0,t_f],\mathcal{H})$ and $\boldsymbol{P}(t) \in C([0,t_f],\mathcal{L}(\mathcal{H}))$ such that $\hat{\boldsymbol{z}}_P(t)$ solves (11) and $\boldsymbol{P}(t)$ satisfies the Riccati equation (5), coupled to (6) and (10).

The other main result concerns the error dynamics.

$$\boldsymbol{\phi}(\boldsymbol{e},t) = \boldsymbol{F}(\boldsymbol{z},t) - \boldsymbol{F}(\boldsymbol{z}-\boldsymbol{e},t) - \boldsymbol{D}F(\boldsymbol{z}-\boldsymbol{e},t)(\boldsymbol{e}). \tag{12}$$

The error $\boldsymbol{e}(t) = \boldsymbol{z}(t) - \hat{\boldsymbol{z}}_P(t)$ between the system state $\boldsymbol{z}(t)$ and the observer state $\hat{\boldsymbol{z}}_P(t)$ has the dynamics

$$\frac{\partial \boldsymbol{e}(t)}{\partial t} = \boldsymbol{A}\boldsymbol{e}(t) - L_P(t)\boldsymbol{C}\boldsymbol{e}(t) + \boldsymbol{DF}(\boldsymbol{z}(t)-\boldsymbol{e}(t),t)\boldsymbol{e}(t) + \boldsymbol{\phi}(\boldsymbol{e}(t)) - L_P(t)\boldsymbol{\eta}(t) - \boldsymbol{G}\boldsymbol{w}(t). \tag{13}$$

These dynamics are well-posed, and furthermore, locally exponentially stable.

Theorem 5. Let the system $(\boldsymbol{A}_d(t),\boldsymbol{C})$, where $\boldsymbol{A}_d(t) = \boldsymbol{A} + \alpha\boldsymbol{I} + \boldsymbol{DF}(\hat{\boldsymbol{z}}_P(t),t)$, be uniformly detectable, and $(\boldsymbol{A}_d(t),\boldsymbol{W}^{1/2}(t))$ be uniformly stabilizable. Assume that there exist time $T > 0$ and positive numbers $m > 1$, $\epsilon_\varphi > 0$ and $\delta_\varphi > 0$ such that if $\|\boldsymbol{e}\|_{\mathcal{H}} \le \epsilon_\phi$, then defining $\boldsymbol{\phi}(\cdot)$ as in (12),

$$\|\boldsymbol{\phi}(\boldsymbol{e},t)\|_{\mathcal{H}} \le \delta_\phi \|\boldsymbol{e}\|_{\mathcal{H}}^m, \quad 0 \le t \le T. \tag{14}$$

In the absence of disturbances, $\boldsymbol{\xi} = \boldsymbol{\eta} \equiv 0$, there exists ϵ, $\delta_{e,0} > 0$, $M_e > 0$, such that if $\|e(0)\| = \|\boldsymbol{z}(0) - \hat{\boldsymbol{z}}_P(0)\|_{\mathcal{H}} < \epsilon$ then for $t \ge 0$,

$$\|\boldsymbol{z}(t) - \hat{\boldsymbol{z}}_P(t)\|_{\mathcal{H}} \le M_e \exp(-\alpha_e(t-0))\|\boldsymbol{z}(0) - \hat{\boldsymbol{z}}_P(0)\|_{\mathcal{H}}.$$

If disturbances are present, the estimation error is bounded.

Corollary 6. Consider the same assumptions as in Theorem 5 except that there are disturbances $\boldsymbol{\xi}(t),\boldsymbol{\eta}(t)$ satisfying for some $\delta_d > 0$ $\|\boldsymbol{\xi}(t)\|, \|\boldsymbol{\eta}(t)\| \le \delta_d$. There exists $\delta_e > 0$ such that if $\|\boldsymbol{e}(0)\| \le \delta_e$ and, the estimation error is bounded.

These results will be illustrated by implementation of an EKF observer for a magnetic drug delivery system modelled by a system of semilinear partial differential equations.

REFERENCES

Afshar, S., Morris, K.A., and Khajepour, A. (2018). State-of-charge estimation using an EKF-based adaptive observer. *IEEE Transactions on Control Systems Technology*.

Afshar, S., Germ, F., and Morris, K. (2020). Well-posedness of extended Kalman filter equations for semilinear infinite-dimensional systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 1210–1215. IEEE.

Ahrens, J.H. and Khalil, H.K. (2007). Closed-loop behavior of a class of nonlinear systems under EKF-based control. *IEEE Trans. Automat. Control*, 52(3), 536–540. doi:10.1109/TAC.2007.892376.

Alonge, F., Cangemi, T., D'Ippolito, F., Fagiolini, A., and Sferlazza, A. (2014). Convergence analysis of extended Kalman filter for sensorless control of induction motor. *IEEE Transactions on Industrial Electronics*, 62(4), 2341–2352.

Baras, J., Bensoussan, A., and James, M. (1988). Dynamic observers as asymptotic limits of recursive filters: Special cases. *SIAM Journal on Applied Mathematics*, 48(5), 1147–1158.

Bitzer, M. and Zeitz, M. (2002). Design of a nonlinear distributed parameter observer for a pressure swing adsorption plant. *Journal of process control*, 12(4), 533–543.

Curtain, R.F. (1975). A survey of infinite-dimensional filtering. *SIAM Review*, 17(3), 395–411.

Curtain, R.F. and Pritchard, A.J. (1978). *Infinite dimensional linear systems theory*, volume 8 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Berlin-New York.

Curtain, R. and Pritchard, A. (1976). The infinite-dimensional Riccati equation for systems defined by evolution operators. *SIAM Journal on Control and Optimization*, 14(5), 951–983.

Curtain, R.F. (1976). Estimation theory for abstract evolution equations excited by general white noise processes. *SIAM Journal on Control and Optimization*, 14(6), 1124–1150.

Efe*, M., Özbay, H., and Samimy, M. (2005). Infinite dimensional and reduced order observers for Burgers equation. *International Journal of Control*, 78(11), 864–874.

Einicke, G.A. and White, L.B. (1999). Robust extended Kalman filtering. *IEEE Transactions on Signal Processing*, 47(9), 2596–2599.

Elizabeth, S. and Jothilakshmi, R. (2015). Convergence analysis of extended Kalman filter in a noisy environment hrough difference equations. *International Journal of Differential Equations and Applications*, 14(2).

Germ, F. (2019). *Estimation for Linear and Semi-linear Infinite-dimensional Systems*. Master's thesis, University of Waterloo.

Grewal, M.S. and Andrews, A.P. (2011). *Kalman filtering: theory and practice using MATLAB*. John Wiley & Sons.

Kai, X., Liangdong, L., and Yiwu, L. (2011). Robust extended Kalman filtering for nonlinear systems with multiplicative noises. *Optimal Control Applications and Methods*, 32(1), 47–63.

Kai, X., Wei, C., and Liu, L. (2010). Robust extended Kalman filtering for nonlinear systems with stochastic uncertainties. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 40(2), 399–405.

Liang, D.F. (1983). Comparisons of nonlinear recursive filters for systems with nonnegligible nonlinearities. In *Control and Dynamic Systems*, volume 20, 341–401. Elsevier.

Miranda, R., Moreno, J., Chairez, J., and Fridman, L. (2012). Observer design for a class of hyperbolic PDE equation based on a distributed super twisting algorithm. In *12th International Workshop on Variable Structure Systems (VSS)*, 367–372. IEEE.

Reif, K., Günther, S., Yaz, E., and Unbehauen, R. (1999). Stochastic stability of the discrete-time extended Kalman filter. *IEEE Transactions on Automatic Control*, 44.

Reif, K., Günther, S., Yaz, E., and Unbehauen, R. (2000). Stochastic stability of the continuous-time extended Kalman filter. *IEEE Proceedings in Control Theory and Applications*.

Reif, K., Sonnemann, F., and Unbehauen, R. (1998). An EKF-based nonlinear observer with a prescribed degree of stability. *Automatica*, 34(9), 1119–1123.

Reif, K. and Unbehauen, R. (1999). The extended Kalman filter as an exponential observer for nonlinear systems. *IEEE Transactions on Signal Processing*, 47(8), 2324–2328.

Ribeiro, M.I. (2004). Kalman and extended Kalman filters: Concept, derivation and properties. *Institute for Systems and Robotics*, 43, 46.

Rigatos, G., Siano, P., Melkikh, A., and Zervos, N. (2017). Highway traffic estimation using nonlinear Kalman filtering. *Intell Industrial Systems*.

Simon, D. (2006). *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons.

Wu, X., Jacob, B., and Elbern, H. (2015). Optimal control and observation locations for time-varying systems on a finite-time horizon. *SIAM Jour. Control and Optim.*, 54(1), 291–316.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Dynamic Consensus of Nonlinear Oscillators under Weak Coupling ⋆

Anes Lazri * **Elena Panteley** ** **Antonio Loría** **

* *Laboratoire des signaux et systèmes, Univ-Paris Saclay, CentraleSupélec, France*
** *Laboratoire des signaux et systèmes, CNRS, France*

Abstract: Dynamic consensus is a property of networked systems that pertains to the case in which all the interconnected systems synchronise their motions and a collective behaviour arises. If the coupling strength is large such behaviour may be modelled by a single system, but if it is weak, the behaviour is best modelled by a reduced-order network. For networks of homogeneous Stuart-Landau oscillators under weak coupling, we characterise the dimension and dynamics of such reduced-order network in function of the coupling strength.

Keywords: Networked Control Systems, Nonlinear Systems and Control, Stability
AMS subject classification: 34D06, 93D20

1. CONTEXT

We analyse the dynamical behaviour of N identical Stuart-Landau oscillators interconnected via a distributed consensus-control law and with interconnection gain $\gamma > 0$,

$$\dot{z}_j = f(z_j) + \mu z_j - \gamma \sum_{k=1}^{N} a_{kj}(z_j - z_k), \qquad j \in \{1, 2, ..., N\}, \tag{1}$$

where $z_j, \mu \in \mathbb{C}$, and $f : \mathbb{C} \to \mathbb{C}$ is defined as

$$f(z_j) = -z_j|z_j|^2. \tag{2}$$

Equations of interconnected Stuart-Landau systems, such as (1)–(2), are often used as a universal dynamical system to model networks exhibiting oscillations, such as lasers, genetic and neuronal networks, among others (Hasty et al., 2001; Soriano et al., 2013).

More precisely, we are interested in the possible synchronised behaviour of these oscillators under the effect of the last term on the right-hand-side of (1). Two factors affect the collective behaviour of the multi-agent system. On one hand, the network's topology, which is defined by the coefficients a_{kj}, and on the other hand, the magnitude of the coupling strength. For instance, for networks of identical oscillators with an underlying undirected connected-graph topology and with $\gamma > 0$ sufficiently large, the networked systems trajectories $z_j(t)$ converge to the solution of an averaged dynamical system,

$$\dot{z}_m = F_m(z_m, e) \tag{3}$$

where $z_m = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} z_k$ and e is a synchronisation error defined as

$$e := z - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top z \quad \Longleftrightarrow \quad \begin{bmatrix} z_1 - z_m \\ \vdots \\ z_N - z_m \end{bmatrix} \tag{4}$$

—see (Pogromsky et al., 1999). For networks of heterogeneous oscillators, *i.e.*, with different μ_js in (1), in (Maghenem et al., 2016) it is shown that for sufficiently large values of γ the synchronisation error defined in (4) is ultimately bounded and the solutions are frequency-synchronised. This is significant because the function $(z_m, e) \mapsto F_m(z_m, e)$ is such that

$$F_m(z_m, 0) = -z_m|z_m|^2 + \mu_m z_m \tag{5}$$

where $\mu_m \in \mathbb{C}$ is defined as $\mu := \mu_R + i\mu_I$ and

$$\mu_R := \frac{1}{\sqrt{N}} \sum_{1}^{N} \mu_{Rk}, \quad \mu_I := \frac{1}{\sqrt{N}} \sum_{1}^{N} \mu_{Ik}. \tag{6}$$

In other words, for sufficiently large values of the coupling strength γ the response of each oscillator in (1) approaches that of a *single* emergent oscillator of the same nature,

$$\dot{z}_m = -z_m|z_m|^2 + \mu_m z_m, \quad z_m \in \mathbb{C}. \tag{7}$$

In general, for heterogeneous systems, such motion is called dynamic consensus (Panteley and Loría, 2017). For Eq. (7), the solutions are either periodic trajectories whose frequency and amplitude are defined by μ or the unstable equilibrium point $\{z_m = 0\}$. Because the right-hand side of (7) corresponds to $F_m(z_m, 0)$, this equation describes the asymptotic collective behaviour of the networked systems, provided they enter in synchrony.

Besides the coupling gain being relatively high, the fact that the collective behaviour of the networked systems (1) may be approximated asymptotically by that of a single oscillator is also a consequence of the graph being connected and undirected. Indeed, in this case, the associated Laplacian matrix

$$L_N := [\ell_{kj}] \in \mathbb{R}^{N\times N}, \qquad \ell_{kj} = \begin{cases} \sum_{l\in\mathcal{N}_k} a_{kl} & k = j \\ -a_{kj} & k \neq j, \end{cases} \tag{8}$$

has exactly one null eigenvalue and $v_1 := \frac{1}{\sqrt{N}}\mathbf{1}_N$ is its associated left eigenvector. That is,

$$z_m := v_1^\top z \quad \text{and} \quad e = [I_N - v_1 v_1^\top]z. \tag{9}$$

⋆ This work was supported by the French ANR via project HANDY, contract number ANR-18-CE40-0010 and by CEFIPRA under the grant number 6001-A.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

It is well-documented in the literature that for relatively small values of γ networks of Stuart-Landau oscillators can exhibit a rich variety of behaviours, such as chaotic motion, emergence of cluster states, and coherence resonance, among others (Golubitsky et al., 2012). Such richer behaviours cannot be captured by that of a single oscillator with dynamics as in (7). In (Tumash et al., 2019) it is shown that, in some cases, rich behaviour can be characterised by a *network* of reduced order. Specifically, the main results in (Tumash et al., 2019) only apply to networks of even dimension and a specific topology, and the reduced-order model's interconnections are nonlinear.

In this extended abstract we consider networks of arbitrary dimension and with circulant-graph topology. We show that the dimension of the reduced model depends on the magnitude of the coupling strength γ relative to the eigenvalues of the Laplacian matrix L_N. Significantly, and in contrast to (Tumash et al., 2019), the reduced-order model has exactly the same structure as the original multiagent system, with linear interconnections.

2. MAIN RESULT

Consider systems modelled by (1), which we rewrite in the compact form

$$\dot{z} = F(z) + \gamma \tilde{L}_N z \tag{10}$$

where $z = [z_1 \ z_2 \ \cdots \ z_N]^\top$,

$$F(z) = [\, f(z_1) \quad f(z_2) \quad \cdots \quad f(z_N) \,]^\top$$

and

$$\tilde{L}_N := \Big[-L_N + \frac{\mu}{\gamma} I_N \Big], \tag{11}$$

under the following hypothesis.

Assumption 1. The network is undirected and connected, has an underlying circulant-graph topology. That is, the adjacency matrix is circulant.

Assumption 1 implies that the Laplacian L_N with coefficients defined in (8) is a circulant matrix. That is, the $(k+1)$st row of L_N corresponds to the kth row in which the last element, ℓ_{kN}, is placed first in the $(k+1)$st row and all other elements are shifted right. An example of a network satisfying Assumption 1 corresponds to one with an underlying ring topology, and in which the nodes may have supplementary cross-links, but with the restriction that each node has the same number of neighbours, $m \geq 2$. See Fig. 1 in Section 3 for an illustration.

Then, we have the following.

Proposition 2. (Main result). Consider a network of N Stuart-Landau oscillators with dynamics (10)–(11), such that the Laplacian L_N satisfies Assumption 1. Then, there exists $\gamma_m > 0$ such that, for each $\gamma > \gamma_m$, there exists $N_R(\gamma) < N$ and a network of reduced order $N_R(\gamma)$, with dynamics given by

$$\dot{z}_R = F(z_R) + \gamma \tilde{L}_R z_R, \tag{12}$$

where $\tilde{L}_R \in \mathbb{C}^{N_R \times N_R}$, $z_R \in \mathbb{C}^{N_R}$, $z = [z_{R_1} \ z_{R_2} \ \cdots \ z_{R_{N_R}}]^\top$,

$$F(z) = [\, f(z_{R_1}) \quad f(z_{R_2}) \quad \cdots \quad f(z_{R_{N_R}}) \,]^\top$$

Moreover, the solutions of (10) satisfy

$$\lim_{t\to\infty} z(t) - \bar{z}_R(t) = 0 \tag{13}$$

for any initial conditions in $\mathbb{C}^N$, where $\bar{z}_R = M z_R$ and $M \in \mathbb{C}^{N \times N_R}$ is a matrix such that its columns generate a subspace of dimension N_R. □

In this extended abstract we do not provide a complete proof of Proposition 2, but the main rationale behind.

First, we observe that because the graph is connected (see Assumption 1), L_N has a unique zero eigenvalue and it admits the Jordan decomposition

$$L_N = U \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_2 \end{bmatrix} U^*, \tag{14}$$

where $U^* \in \mathbb{C}^{N\times N}$ denotes the conjugate transpose of U, which is orthonormal, so $UU^* = U^*U = I_N$. Furthermore, $\Lambda_2 \in \mathbb{C}^{N-1\times N-1}$ is a diagonal matrix whose elements correspond to the nonzero eigenvalues of L_N. Then, in view of its definition—see Eq. (11), $\tilde{L}_N$ satisfies the same decomposition as L_N. That is, denoting by $\lambda_i(L_N)$ the eigenvalues of L_N, after (11), we have

$$\tilde{L}_N = U \begin{bmatrix} \frac{\mu}{\gamma} & 0 & \cdots & & 0 \\ 0 & -\lambda_2(L_N) + \frac{\mu}{\gamma} & & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & & \cdots \ 0 & -\lambda_N(L_N) + \frac{\mu}{\gamma} \end{bmatrix} U^*. \tag{15}$$

Clearly, the number of eigenvalues of the matrix above with positive real part varies from one to N depending on the value of γ. On the other hand, by convention, the eigenvalues λ_i are ordered in a way that $0 < \Re e\{\lambda_2\} \leq \Re e\{\lambda_3\} \leq \ldots \leq \Re e\{\lambda_N\}$.

Furthermore, under Assumption 1, if N is odd, the nonzero eigenvalues λ_k come in conjugate pairs, that is $\Re e\{\lambda_k\} = \Re e\{\lambda_{k+1}\}$ for all $k \in \{2, 4, \ldots, N-1\}$. Hence, for any γ such that for some $k \in \{2, 4, \ldots, N-3\}$

$$\Re e\{\mu/\lambda_{k+2}\} < \gamma < \Re e\{\mu/\lambda_k\}$$

the matrix in (15) has necessarily an odd number $N_R(\gamma) = k+1$ eigenvalues with positive real part, and $N_R(\gamma) = N$ for any $\gamma < \Re e\{\mu/\lambda_{N-1}\}$. Thus, for any $\gamma > \gamma_m := \Re e\{\mu/\lambda_N\} = \Re e\{\mu/\lambda_{N-1}\}$, the matrix in (15) has $N_R(\gamma) < N$ eigenvalues with positive real part. Similar arguments apply to the case in which N is even, considering that $\lambda_N \in \mathbb{R}_{>0}$.

Now, for the sake of argument, let us disregard momentarily the nonlinear terms in (10), $F(z)$. For the system $\dot{z} = \gamma \tilde{L}_N z$, the eigenvalues with positive real part in $\tilde{L}_N$ generate unstable modes while those with negative real part generate stable ones. That is, the solution may be written as

$$z(t) = v_1 v_1^* z(t) + v_2 v_2^* z(t) + \cdots + v_{N_R} v_{N_R}^* z(t) + e(t), \tag{16}$$

where $v_k \in \mathbb{C}^N$, for all $k \in \{1, 2, \ldots, N_R\}$ are eigenvectors associated with eigenvalues with positive real part and $e(t)$ contains the contributions to the solution generated by the stable modes. As $e(t) \to 0$ only the contributions of the unstable modes remain. This motivates the choice for the synchronisation errors,

$$e := z - U_1 U_1^* z, \tag{17}$$

where $U_1 \in \mathbb{C}^{N\times N_R}$ is such that its columns correspond to the N_R eigenvectors associated with the N_R eigenvalues with positive real part, that is, v_k farther above. Since v_k denote the columns of U_1, note that if $N_R = 1$ we recover the expression (4). On the other hand, on the synchronisation manifold $\{e = 0\}$ the dynamics of the

system (10) reduces to that of reduced-order network evoked in Proposition 2. Exponential stability of $\{e = 0\}$ can be established along the lines of (Tumash et al., 2019).

Now, the statement of Proposition 2 asserts that the solution of the networked system converges asymptotically to a trajectory that we denote $\bar{z}_R(t)$. This trajectory, loosely speaking, may be regarded as a linear combination of the elements, $z_{R_k}(t)$, of the vector $z_R(t)$, solution of (12). Indeed, introducing the notation $M = [v'_1\ v'_2\ \cdots\ v'_{N_R}]$, we see that

$$\bar{z}_R = v'_1 z_{R_1} + v'_2 z_{R_2} + \cdots + v'_{N_R} z_{R_{N_R}} \tag{18}$$

and the N_R vectors v'_k are uniquely determined by the properties of the Laplacian L_N and the interconnection strength γ. It is important to remark that $v'_k \neq v_k$. As a matter of fact, it may be shown that

$$M := U_1[\,U_{11}^* - U_{21}^*(U_{22}^*)^{-1}U_{12}^*\,],$$

where the matrices on the right-hand side above come from a suitable partition of U and its conjugate transpose,

$$U =: \begin{bmatrix} U_{11} & U_{21} \\ U_{12} & U_{22} \end{bmatrix}, \quad U^* = \begin{bmatrix} U_{11}^* & U_{21}^* \\ U_{12}^* & U_{22}^* \end{bmatrix}. \tag{19}$$

With these definitions, it may also be shown that $e = z - Mz_R$ or that Eq. (16) is equivalent to $z(t) = e(t) + Mz_R(t)$ so, in the limit, as $e(t) \to 0$, $z(t) = Mz_R(t)$—cf. (13). In that regard, (12) is reminiscent of the zero-dynamics of the networked system (10) with respect to the converging output $e \in \mathbb{C}^N$. This corroborates the initial observation that, for complex networked nonlinear systems, the collective behaviour in the state of synchronisation, that is, on the manifold $\{e = 0\}$ depends on the coupling strength γ.

The importance of the latter observations transcend the rationale behind the proof of Proposition 2, which is omitted due to space constraints. The fact that, asymptotically, $z(t) = Mz_R(t)$, implies that the solutions of the (potentially) high-order network (10) may be reconstructed using the solutions of a reduced-order model, at least asymptotically, as synchronisation is reached. This is possible using a change of coordinates that leads to (12).

3. EXAMPLE

We have that if the coupling gain is large ($\gamma > \gamma_M$) then, $N_R = 1$; that is, (12) becomes, simply, Eq. (7). In this case, the oscillators synchronise globally. If γ is relatively small ($\gamma_M > \gamma > \gamma_m$), which is the case of most interest here, there exists a network of reduced order $N > N_R(\gamma) > 1$.

For illustration, let us consider a network of eleven oscillators modelled as in (1), with $\mu = 1+2i$ and interconnected over a graph satisfying Assumption 1, which is illustrated in Fig. 1. We consider two cases that pertain to different values of the coupling strength. First, we set $\gamma < 1$, which is relatively small given that $\mu_R = 1$. Indeed, for such γ, $\tilde{L}_N$ in (11) has three eigenvalues with positive real part. Hence, according to Proposition 2, there exists a reduced-order network of dimension $N_R = 3$ whose behaviour approaches asymptotically that of the original one. Then, the coupling strength is increased to $\gamma = 2.5$, so only the first element in the diagonal of the matrix in (15) has positive real part. Consequently, it is expected that the collective behaviour of the networked systems approach asymptotically that of a single one with dynamics as in Eq. (7).

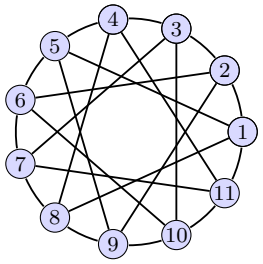


Fig. 1. Example of a ring graph in which each node has the same (even) number of undirected links

In both simulation tests the initial conditions are set to

$$\begin{cases} z_j(0) = 1 + i & \forall\, j \in \{1, 2, \ldots, 5\}, \\ z_6(0) = 0, \\ z_j(0) = -1 - i & \forall\, j \in \{7, 8, \ldots, 11\}. \end{cases} \tag{20}$$

The results are shown in Figs. 2–6, for a coupling strength set to $\gamma = 0.75$, and in Fig. 7 for $\gamma = 2.5$.

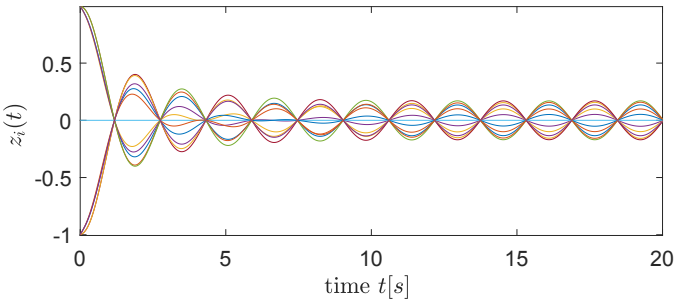


Fig. 2. Response of $N = 11$ oscillators under relatively small coupling strength $\gamma = 0.75$

In Fig. 2 are shown the trajectories of all oscillators. It may be appreciated that they all synchronise in frequency, but the oscillations have different amplitudes and there are two groups of in-phase oscillators. This is also appreciated from Fig. 3, which shows the evolution of the synchronisation errors as defined in (17) and from the plot on the left in Fig. 4 where the trajectories are depicted on the complex plane.

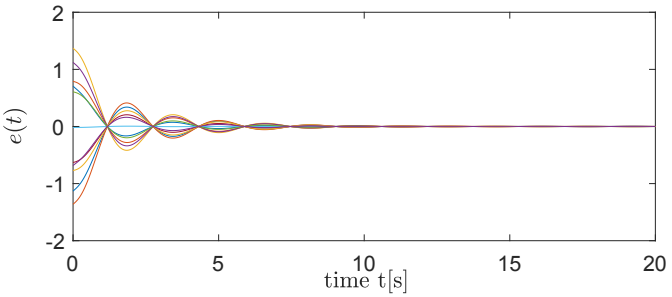


Fig. 3. Synchronisation error $e(t)$ with e as in (17)

The behaviour of the three oscillators composing the reduced-order network (12) is depicted in Fig. 5.

Three different behaviours appear: one, generated by the mode related by the null eigenvalue of L_N, is equivalently equal to 0 since the initial condition is set at the origin, which is an equilibrium. Two other modes, related to the conjugate eigenvectors v_1 and v_2, generate trajectories which are opposite to one another.

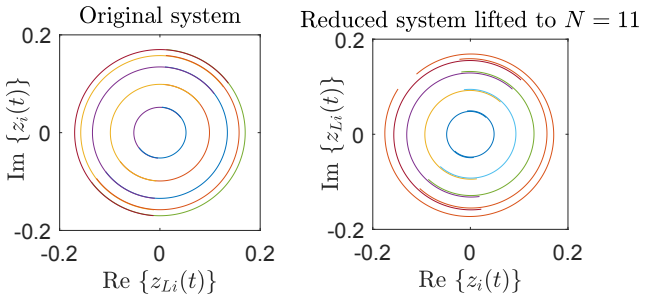


Fig. 4. Oscillators' trajectories on the complex plane for the network of 11 oscillators (left plot) and for their behaviour reconstituted from that of the reduced-order network (right plot)

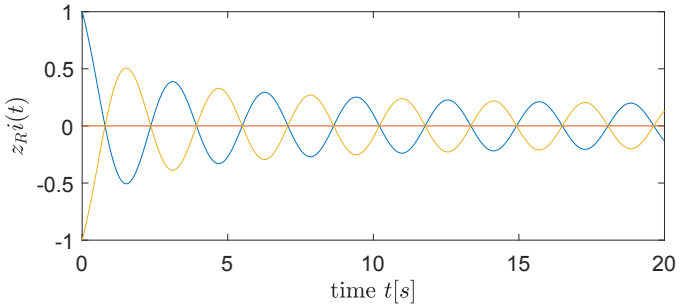


Fig. 5. Simulations results of $N_R = 3$ reduced order system

Note that the oscillators of the original network of dimension 11 do not necessarily synchronise their behaviour with any of the three oscillators described by (12). However, the behaviour of each $z_i(t)$ in (1) may be reconstructed, asymptotically, using the responses of the reduced-order network. The reconstituted behaviour, against time, is shown in Fig. 6 and on the complex plane in the right plot on Figure 4.

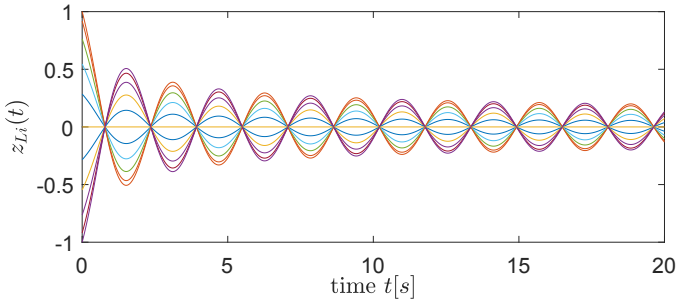


Fig. 6. Simulations results for the reduced system lifted to $N = 11$

In Fig. 7 are shown the oscillators' responses with the initial conditions as in (20) and with high coupling strength, $\gamma = 2.5$. In this case, all the oscillators' trajectories converge exponentially to zero. This is explained by the fact that they approach the dynamics of the average system in (7) with zero initial condition $z_m(0) = 0$, which corresponds to the average of $z_i(0)$ in (20).

Finally, to illustrate the influence of the initial conditions, in Fig. 8 we show the response of the oscillators, with coupling gain $\gamma = 2.5$ and in which case, $z_6(0) = 1 + i$. Asymptotically, all the systems converge to the trajectory of one averaged oscillator, thereby achieving dynamic consensus. In addition, the asymptotic behaviour of the network may also be reconstructed from that of the averaged motion, from $t \approx 1$s.

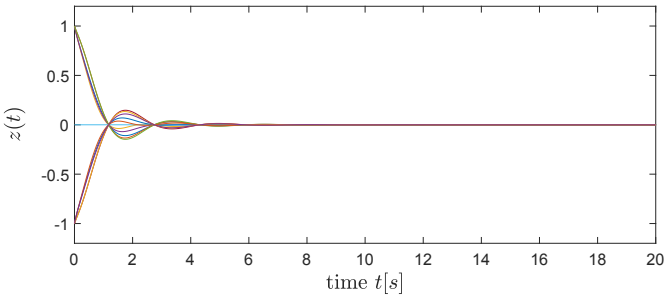


Fig. 7. Oscillators response with $\gamma = 2.5$ with initial conditions as in (20)

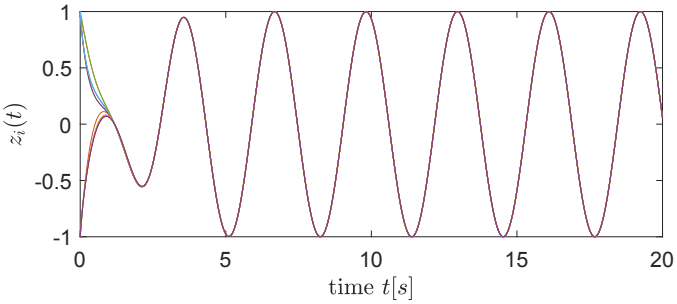


Fig. 8. Oscillators response with $\gamma = 2.5$ with initial conditions as in (20), except for $z_6(0) = -1$

4. CLOSING REMARKS

Our results illustrate how to reconstruct the asymptotic behaviour of complex networked systems, with specific topology, via a model-reduction. Undergoing research is aimed at analysing the reduced-order model and considering other network topologies.

REFERENCES

Golubitsky, M., Stewart, I., and Schaeffer, D.G. (2012). *Singularities and Groups in Bifurcation Theory: Volume II*. Springer Science & Business Media.

Hasty, J., Mcmillen, D., Isaacs, F., and Collins, J. (2001). Computational studies of gene regulatory networks: In numero molecular biology. *Nature reviews. Genetics*, 2, 268–279.

Maghenem, M., Panteley, E., and Loria, A. (2016). Singular-perturbations-based analysis of synchronization in heterogeneous networks: A case-study. In *55th IEEE Conference on Decision and Control*, 2581–2586. Las Vegas, NV, USA.

Panteley, E. and Loría, A. (2017). Synchronization and dynamic consensus of heterogeneous networked systems. *IEEE Trans. on Automatic Control*, 62(8), 3758–3773.

Pogromsky, A., Glad, T., and Nijmeijer, H. (1999). On diffusion driven oscillations in coupled dynamical systems. *International Journal of Bifurcation and Chaos*, 9(4), 629–644.

Soriano, M., Garcia-Ojalvo, J., Mirasso, C., and Fischer, I. (2013). Complex photonics: Dynamics and applications of delay-coupled semiconductors lasers. *Review of Modern Physics*, 85(1), 421–470.

Tumash, L., Panteley, E., Zakharova, A., and Schöll, E. (2019). Synchronization patterns in Stuart–Landau networks: a reduced system approach. *The European Physical Journal B: Condensed Matter and Complex Systems*, 92(5).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Dissipativity in Analysis of Neural Networks

Johan Grönqvist * **Anders Rantzer** **

* *Automatic Control LTH, Lund University, Box 118, SE 22100 Lund, Sweden (e-mail:johan.gronqvist@control.lth.se).*
** *Automatic Control LTH, Lund University, Box 118, SE 22100 Lund, Sweden (e-mail:rantzer@control.lth.se).*

Abstract: Building on the strong connection between dissipativity theory and Integral Quadratic Constraints, we show how feedback loops involving neural networks can be analysed computationally with respect to both stability and robustness. A basic building block is the ReLU (Rectified Linear Unit) nonlinearity and we present both old and new dissipation inequalities that are useful for its analysis.

Keywords: Neural networks, robustness analysis, relaxations, stability of nonlinear systems, convex optimization.

1. INTRODUCTION

The pioneering work on dissipativity in Willems (1972) has been a cornerstone of systems analysis ever since. Many extensions were developed over the years and more recently the concept Integral Quadratic Constraint (IQC) Megretski and Rantzer (1997) emerged in an effort to unify dissipativity analysis with frequency domain methods dating back to Yakubovich (1967). Many IQCs can be expressed as dissipativity inequalities with an explicitly given non-negative storage function. For others, the non-negativity is replaced by a more relaxed constraint and the storage function may not be explicitly given. See Megretski et al. (2010).

In recent years, deep learning with neural networks has achieved impressive successes in many fields. Following in the footsteps of this progress, many groups have worked on robustness for systems involving neural networks. Neural networks are well-suited to the IQC formalism and previous work along this line draws on old results. Static properties like reachability were treated in Hu et al. (2020) and Lipschitz constant estimation was carried out in Fazlyab et al. (2019), which can be used as one part in a larger analysis of robustness of the full system. The closed loop system can also be analysed as a whole as done in Yin et al. (2020), and provides guarantees of global stability in the presence of uncertain dynamics. The IQC formalism has also been used in other contexts to provide other kinds of guarantees, and those kinds of guarantees would transfer to systems with neural networks. The limitation is that the IQC-framework assumes all nonlinear elements, as well as all properties, to be expressed in terms of quadratic forms, limiting the number of questions that the framework can answer. The strength, is that guarantees are global and can be verified even in the face of adversarial attacks on the network.

Here, and in Grönqvist and Rantzer (2022), we present a larger class of constraints that can be used for neural networks, than used in the works cited above. We also provide examples showing that a larger library of constraints can enable stricter bounds on the behaviour of neural networks, and that there is a trade-off in the choice of constraints. After introducing neural networks, we describe three different analysis settings, followed by a listing of the available constraints in each analysis setting. Finally, we present two small examples, highlighting the different levels of expressiveness that can be achieved by choosing an analysis setting and a selection of constraints.

$$\zeta_0 \xrightarrow[b_1]{W_1} \xi_1 \xrightarrow{\varphi} \zeta_1 \xrightarrow[b_2]{W_2} \xi_2 \xrightarrow{\varphi} \zeta_2 \to \cdots \to \zeta_N \xrightarrow[b_{N+1}]{W_{N+1}} \zeta_{N+1}$$

Fig. 1. A neural network, with input ζ_0, output $\zeta_{N+1} = W_{N+1}\zeta_N + b_{N+1}$, and a sequence of N hidden layers with affine mappings $\xi_k = W_k\zeta_{k-1} + b_k$ and nonlinear mappings $\zeta_k = \varphi(\xi_k)$, for $k = 1, \ldots, N$.

2. NEURAL NETWORKS

Figure 1 shows a simple feedforward neural network and lists the defining equations. More generally, the network can be a graph with nodes representing either affine relations or nonlinear activation functions. Activation functions are often scalar and applied componentwise.

The internal variables, ξ_k and ζ_k, are vectors, and we may have several neural networks in our closed loop system. We assemble all components from all variables $\xi_1, \ldots, \xi_N$ into a large vector ξ, and all components from $\zeta_1, \ldots, \zeta_N$ into a large vector ζ, and we write

$$\zeta = \varphi(\xi) \tag{1}$$

where we now have a componentwise application of the nonlinear function φ on a vector with a large number of components.

This structure is described as a repeated nonlinearity, and enables us to find a large set of quadratic constraints to help us obtain guarantees for systems with neural networks.

We note that neural network typically contain affine layers, defined by $\xi_k = W_k\zeta_{k-1} + b_k$ that can be cumbersome in some analyses. We discuss how we handle this issue using the relative setting in section 4.

3. FORMS OF CONSTRAINTS

For the IQC formalism, there are two kinds of blocks in a system. One kind includes linear time invariant systems, and all other blocks are characterized by quadratic inequalities that hold for their inputs and outputs. We describe three kinds of constraints here.

Static constraints, hard IQCs and soft IQCs take the forms

$$\begin{aligned} \begin{pmatrix}\xi(t)\\ \zeta(t)\end{pmatrix}^\top G \begin{pmatrix}\xi(t)\\ \zeta(t)\end{pmatrix} \geq 0 \ \text{ for all } t \\ \int_{t=0}^{T} \begin{pmatrix}\xi(t)\\ \zeta(t)\end{pmatrix}^\top G \begin{pmatrix}\xi(t)\\ \zeta(t)\end{pmatrix} dt \geq 0 \ \text{ for all } T \\ \int_{-\infty}^{\infty} \begin{pmatrix}\hat\xi(i\omega)\\ \hat\zeta(i\omega)\end{pmatrix}^\star \hat G(i\omega) \begin{pmatrix}\hat\xi(i\omega)\\ \hat\zeta(i\omega)\end{pmatrix} d\omega \geq 0. \end{aligned} \tag{2}$$

A static quadratic constraint automatically gives a hard IQC and a hard IQC automatically gives a soft IQC, with the same matrix G used in all cases. We typically have constraints on the structure of G, that reflect the quadratic inequalities they are based on.

We will only list IQCs for the continuous time case, but analogous forms for the discrete time case exist, except for the Popov IQC. The constraints are then combined using the S-procedure to provide guarantees for the neural network or closed loop system, as used in our examples, and as discussed at length in Megretski and Rantzer (1997) and Megretski et al. (2010).

3.1 From Static to Dynamic

There are several ways to construct IQCs from quadratic inequalities, beyond the immediate hard and soft IQCs obtained from static inequalities as described above. Nonnegative expressions in ξ and ζ can be combined using nonnegative matrices or convolution kernels to form IQCs. The two special constructions in our lists, the Popov and Zames-Falb IQCs, are well-established results from the literature, and discussed in more detail in, e.g., Megretski et al. (2010); D'Amato et al. (2001); Safonov and Kulkarni (2000).

4. ANALYSIS SETTINGS

We consider three different analysis settings, and they give us different sets of constraints. We refer to them as the absolute, the relative and the incremental setting, respectively.

The absolute setting uses constraints on the mapping from inputs ξ to outputs ζ.

The relative setting considers a known reference input $\xi^{(\mathrm{ref})}$, that often represents a steady state of the dynamics. It considers constraints on the deviations from the reference signals, in both inputs and output, $\Delta\xi = \xi - \xi^{(\mathrm{ref})}$ and $\Delta\zeta = \zeta - \zeta^{(\mathrm{ref})}$. The nonlinearity from $\Delta\xi$ to $\Delta\zeta$ is now no longer repeated, as we have

$$\Delta\zeta = \varphi\left(\Delta\xi + \xi^{(\mathrm{ref})}\right) - \varphi\left(\xi^{(\mathrm{ref})}\right), \tag{3}$$

which depends on the reference value.

For nonzero reference values, we consider instead the mapping from $\frac{\Delta\xi}{\left|\xi^{(\mathrm{ref})}\right|}$ to $\frac{\Delta\zeta}{\left|\xi^{(\mathrm{ref})}\right|}$ and find that for reference values of the same sign, this is a repeated nonlinearity sharing many of the properties of the relu function. As the reference values are constant and known in advance, this simple rescaling of our signals allows us to use most of our IQCs from the absolute setting.

Referring back to the affine relations in the neural network layers, those relations become linear in the relative setting, which enables us to keep using most of our IQCs for netwoks with nonzero bias terms, b_k, in figure 1.

Lastly, in the incremental setting, we have two arbitrary inputs, $\xi^{(1)}$ and $\xi^{(2)}$, and we study the mapping from $\Delta\xi = \xi^{(1)} - \xi^{(2)}$ to $\Delta\zeta = \zeta^{(1)} - \zeta^{(2)} = \varphi(\xi^{(1)}) - \varphi(\xi^{(2)})$, without any further assumptions on $\xi^{(1)}$ and $\xi^{(2)}$. This has two consequences, compared to the incremental setting. Firstly, we cannot use the rescaling trick to get repeated nonlinearities, and, secondly, we no longer have a function, as knowing $\Delta\xi = 1$ is not sufficient to tell us what $\Delta\zeta$ is.

5. ACTIVATION FUNCTIONS

Several activation functions are used in real networks, but many networks contain almost only relu or leakyrelu activations, with a single softmax nonlinearity in the last layer.

We focus on the relu nonlinearity and note that this enables the corresponding analysis of the very common leakyrelu nonlineary, as the latter satisfies

$$\mathrm{leakyrelu}_a(x) = \max(ax, x) = ax + (1-a)\,\mathrm{relu}(x), \tag{4}$$

where a is a constant that is known in advance.

5.1 Relu

For a single scalar relu, $y = \mathrm{relu}(x) = \max(0, x) \in \mathbb{R}$, we have the basic properties

$$\begin{aligned} y = 0 \ \text{ if } x \leq 0 \\ y \geq 0 \ \text{ and } \ y - x \geq 0 \\ (y-x)y = 0. \end{aligned} \tag{5}$$

For a pair of relus, $y_i = \mathrm{relu}(x_i)$ and $y_j = \mathrm{relu}(x_j)$, we have a sector condition, thanks to a rate-limit and to $\mathrm{relu}(0) = 0$,

$$(y_i - y_j)((x_i - x_j) - (y_i - y_j)) \geq 0. \tag{6}$$

The above relations form the basis for our IQCs.

We now discuss a new form of constraint for the repeated relu-nonlinearity. For our pair of relu operations, we have the linear inequalities of equation (5), and any pairwise product is a quadratic inequality, such as, e.g.,

$$(y_i - x_i)(y_j - x_j) \geq 0 \tag{7}$$

We can convolve with a kernel $h(t)$ such that $h(t) \geq 0$ for all t, to obtain the IQC (where a hat denotes a Laplace transform)

$$\int_{-\infty}^{\infty} (\hat{y}_i(i\omega) - \hat{x}_i(i\omega))^\dagger \hat{h}(i\omega)(\hat{y}_j(i\omega) - \hat{x}_j(i\omega))\, d\omega \geq 0. \quad (8)$$

Including all pairwise producs of linear inequalities, and writing it in terms of our vectors ζ and ξ with a large number of components, we have the IQC

$$\int_{-\infty}^{\infty} \begin{bmatrix} \hat{\zeta}(i\omega) - \hat{\xi}(i\omega) \\ \hat{\zeta}(i\omega) \end{bmatrix}^\dagger \hat{h}(i\omega) \begin{bmatrix} \hat{\zeta}(i\omega) - \hat{\xi}(i\omega) \\ \hat{\zeta}(i\omega) \end{bmatrix} d\omega \geq 0. \quad (9)$$

5.2 Other activation functions

While we focus on leakyrelu and relu as the most common activation functions, other nonlinearities exist, and are primarily of two kinds.

The first kind is a repeated nonlinearity, much like the relu-case, but lacking some of the inequalities from section 5.1. They often satisfy a rate-limit, enabling some of our constraints to hold also for this class of activation functions.

The second kind is the softmax nonlinearity, which is not a repeated nonlinearity. Instead, it is a vector valued function, and it is nonvanishing for vanishing input. These two properties limit the set of constraints available.

6. CONSTRAINTS

We list static constraints and IQCs available to us for the repeated relu-nonlinearity. As discussed above, this enables a corresponding set of constraints on the leakyrelu, thanks to equation (4).

This section assumes a repeated relu-nonlineariy whose input is a vector or vector valued signal. These may be chosen as any subset of the relu-units in the system, i.e., of the vectors ξ and ζ from equation 1.

6.1 Absolute

We look for positive semidefinite quadratic forms in the input and output of the repeated relu-unit.

(1) We can use the quadratic equality of equation (5), multiplied by an arbitrary real number, as a static constraint, and we obtain the corresponding hard and soft IQCs.
(2) As discussed in section 5.1, products of the linear inequalities give us static and dynamic constraints when multiplied with nonnegative coefficients or convolved with nonnegative kernels.
(3) The rate-limit inequalities, equation (6), are also quadratic, and we can use them for any pair of relu-units, giving us static and dynamic constraints.
(4) The Popov IQC is a special construct that we can use for our relu-nonlinearity
(5) There have been several works based on ideas of Zames and Falb, and,we use the results from D'Amato et al. (2001), which are formulated as IQCs, but also give us static constraints as a special case if we set the frequency dependent parts to zero.

6.2 Relative

In the relative setting, we have a set of reference values, as discussed in section 4, and we rescale our inputs and outputs into three sets of repeated nonlinearities, as described in that section.

For rate-limits, we use the full set of inputs and outputs, but for some other constructions, we use the three sets as separate sets of repeated nonlinearities.

(1) The relu units still satisfy linear inequalities in the relative setting, and we again multiply them with nonnegative parameters or convolve with nonnegative kernels to obtain a large family of constraints.
(2) The rate-limit still holds, within each of our three sets of repeated nonlinearties, analogous to the absolute setting.
(3) The rate-limit can be used in an additional way in the relative setting, as every signal is a difference between actual signal and reference signal.
(4) Within each of our three sets of repeated nonlinearities, the results of D'Amato holds, and gives us static and dynamic constraints.

6.3 Incremental

In the incremental case, the rate-limit is our only remaining source of inequalities, and only in a single variant. The linear inequalities no longer hold for the difference between arbitrary signals, and the Zames-Falb construction cannot be used, as our nonlinearity is not a function, as mentioned in section 4.

6.4 Other Activation Functions

Many other activation functions satisfy rate-limits, and the rate-limit related constraints hold for these as well. The corresponding IQCs also hold, similar to the case for the relu-nonlinearity.

As described in section 5.2, the set of constraints for the softmax function is restricted in a different way, due to its special structure.

7. SELECTION OF CONSTRAINTS

The library of constraints is provided here with the aim of completeness, but in practical applications, a selection of which constraints to use is often needed. In particular, for deep neural networks, even the most efficient SDP solvers will struggle as the network grows larger. This can be alleviated by selecting constraints that lead to sparse SDP problems, and we have found it useful to include constraints that couple nonlinearites from adjacent blocks in the block-diagram.

In the neural network, this amounts to constraints that mix signals from adjacent layers, and for a more general setup, it will also mix, e.g., the signals in the first layer of the neural network with the nonlinear block before it, if such constraints are available.

This retains sparsity in the problem, and keeps most of the expressiveness of the IQC formalism.

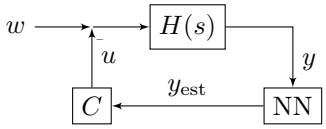


Fig. 2. Block diagram of system used in Closed Loop example.

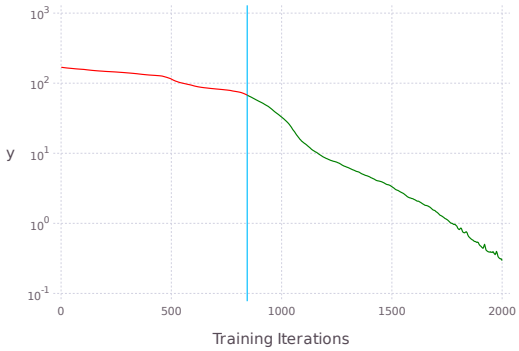


Fig. 3. Training loss vs. number of gradient steps, with guaranteed stability achieved after 884 steps.

8. EXAMPLE

8.1 Static Gain of a Deep Network

We consider a small network defined by

$$\begin{aligned} x_1 &= \begin{pmatrix} 1 \\ -1 \end{pmatrix} y_0; \quad y_1 = \text{relu}(x_1) \\ x_2 &= \begin{pmatrix} 1 & -1 \end{pmatrix} y_1 \end{aligned} \tag{10}$$

where the input y_0 and output x_2 are both scalar.

The network was designed so that $x_2 = y_0$ in all cases, and we will compare analyses of the static gain of this network in the absolute and incremental settings.

In the incremental setting, all we have access to is the rate-limit, and we cannot exclude that $\text{relu}(x) = x$ might hold. In that case, the static gain of the network would be 2, and our analysis is unable to provide a lower bound. In particular, for a deep network that is a repetition of the above structure, the gain bounds of the layers are multiplied together to a gain bound that grows exponentially with the depth of the network.

In the absolute setting, we have a more expressive set of constraints, allowing us to obtain a gain bound of 1, and the gain bound no longer grows exponentially with the depth of the network.

8.2 Closed Loop Stability

We consider a double integrator controlled by a stabilizing PD-controller. The output is filtered through a neural network before passing to the controller. Our network structure is a few-layer variant of the network from the previous example, with the first-layer weights fixed, and with remaining weights trained from random initialization. We use the main theorem of Megretski and Rantzer (1997) together with the KYP lemma discussed in Rantzer (1996) to obtain closed-loop stability guarantees based on the IQC formalism.

Our training progress is shown in figure (3) and sees decreasing loss for a large number of iterations, with our use of the IQC framework giving us guaranteed global stability after 884 gradient steps.

9. CONCLUSION

We provide a list of constraints that can be used for obtaining guarantees for systems involving neural networks. We discuss new constraints, as well as many from previous literature.

We discuss three analysis settings, and provide two small examples showing that we can obtain guarantees using the formalism, and that the absolute setting can provide stronger guarantees than the incremental setting.

ACKNOWLEDGEMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

D'Amato, F., Rotea, M., Megretski, A., and Jönsson, U. (2001). New results for analysis of systems with repeated nonlinearities. *Automatica*, 37(5), 739 – 747.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G.J. (2019). Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Grönqvist, J. and Rantzer, A. (2022). Integral quadratic constraints for neural networks. *European Control Conference*.

Hu, H., Fazlyab, M., Morari, M., and Pappas, G.J. (2020). Reach-sdp: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 5929–5934.

Megretski, A. and Rantzer, A. (1997). System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6), 819–830.

Megretski, A., Jönsson, U., Kao, C.Y., and Rantzer, A. (2010). *Integral Quadratic Constraints*, volume 1. Taylor & Francis.

Rantzer, A. (1996). On the kalman—yakubovich—popov lemma. *Systems & Control Letters*, 28(1), 7–10.

Safonov, M. and Kulkarni, V. (2000). Zames-falb multipliers for mimo nonlinearities. In *Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No.00CH36334)*, volume 6, 4144–4148 vol.6.

Willems, J.C. (1972). Dissipative dynamical systems, part I: General theory; part II: Linear systems with quadratic supply rates. *Arch. Rational Mechanics and Analysis*, 45(5), 321–393.

Yakubovich, V. (1967). Frequency conditions for the absolute stability of control systems with several nonlinear or linear nonstationary units. *Autom. Telemech.*, 5–30.

Yin, H., Seiler, P., and Arcak, M. (2020). Stability analysis using quadratic constraints for systems with neural network controllers.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A multilevel fast-marching method

Marianne Akian, Stéphane Gaubert * **Shanqing Liu** **

* *Inria and CMAP, École polytechnique, IP Paris, CNRS, 91128 Palaiseau Cedex, France (e-mail: Firstname.Name@inria.fr)*
** *CMAP, École polytechnique, IP Paris, CNRS, and Inria, 91128 Palaiseau Cedex, France (e-mail: Shanqing.Liu@polytechnique.edu)*

Abstract: We introduce a new numerical method to approximate the solutions of a class of static Hamilton-Jacobi-Bellman equations arising from minimum time optimal control problems. We rely on several grid approximations, and look for the optimal trajectories by using the coarse grid approximations to reduce the search space for the optimal trajectories in fine grids. This may be thought of as an infinite dimensional version, for PDE, of the "highway hierarchy" method which has been developed to solve discrete shortest path problems. We obtain, for each level, an approximate value function on a sub-domain of the state space. We show that the sequence obtained in this way does converge to the viscosity solution of the HJB equation. Moreover, the number of arithmetic operations that we need to obtain an error of $O(\varepsilon)$ is bounded by $\tilde{O}(1/\varepsilon^{\frac{2d}{1+\beta}})$, to be compared with $\tilde{O}(1/\varepsilon^{2d})$ for ordinary grid-based methods. Here $\beta \in (0,1]$ depends on the "stiffness" of the value function around optimal trajectories, and the notation $\tilde{O}$ ignores logarithmic factors. Under a regularity condition on the dynamics, we obtain a bound of $\tilde{O}(1/\varepsilon^{(1-\beta)d})$ operations, for $\beta < 1$, and this bound becomes $O(|\log\varepsilon|)$ for $\beta = 1$. This allowed us to solve HJB PDE of eikonal type up to dimension 7.

Keywords: Minimum Time, Eikonal equation, Fast Marching Method, Highway Hierarchies.

1. INTRODUCTION

We consider the minimal time optimal control problem, consisting in finding a trajectory minimizing the travel time between two points. The minimal time, together with optimal trajectories, can be obtained by solving a Hamilton-Jacobi-Bellman (HJB) partial differential equation (PDE) of eikonal type. Such an equation is obtained from the dynamic programming principle, and has to be thought in viscosity sense, see Fleming and Soner (2006).

One of the most famous methods to solve the eikonal PDE is the fast-marching method, see Sethian (1996). Though it is computationally efficient, it is still a grid-based method, and hence suffers from the "curse of dimensionality" of the dynamical programming approach. Our algorithm intends to bypass this difficulty by narrowing the search space around optimal trajectories, hence solving a series of PDE, in which the domain is becoming smaller and smaller with the discretize grid becoming finer and finer. This is inspired by the recent development of the "Highway Hierachies" algorithm, which applies to shortest path problems in discrete time and in discrete space, see Sanders and Schultes (2012).

The idea of our algorithm is as follows: instead of computing the optimal trajectories directly in the whole domain, we first find a subdomain which contains the optimal trajectories, then do the search in such a subdomain. This will be achieved by using coarse and fine grids discretization method. We first use a coarse grid to discretize the whole domain, and do a partial fast marching search in this coarse grid. Then, we find a subdomain which contains the true optimal trajectories, by using the approximate value function on the coarse grid, together with the error bound. Finally we use a finer grid to discretize the subdomain, and perform a fast marching search on this fine grid. We repeat this operation, considering different levels of finer and finer grids.

We show that using our algorithm, the final approximation error is as good as the one obtained by directly discretizing the whole domain with the finest grid. Moreover, the number of elementary operations and the size of the memory needed to get an error of ε are considerably reduced. Indeed, let $\kappa(M) := M \log M$, and recall that the fast marching method implemented in a d-dimensional grid with M points requires a number of arithmetic operations of order $\kappa(M)K_d$, in which the constant $K_d \in [2d, 2^d]$ depends on the stencil of the discretization used for the fast marching method. For our method, the number of arithmetic operations is in the order of $\kappa((\frac{1}{\varepsilon})^{\frac{2d}{1+\beta}})C^d$, where $C > 1$ and $\beta \in (0,1]$ are fixed constants (see Section 5), to be compared with $\kappa((\frac{1}{\varepsilon})^{2d})K_d$ for conventional grid-based methods. Moreover, under a regularity condition on the dynamics, our complexity bound reduces to an order of $\kappa((\frac{1}{\varepsilon})^{(1-\beta)d})C^d$ for $\beta < 1$, and $\kappa(d|\log\varepsilon|)$ for $\beta = 1$.

2. HAMILTON-JACOBI EQUATION FOR MINIMUM TIME PROBLEM

Let Ω be an open, bounded domain in $\mathbb{R}^n$. Let S_1 be the unit sphere in $\mathbb{R}^n$, i.e., $S_1 = \{x \in \mathbb{R}^n, \|x\| = 1\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $\mathcal{A} = \{\alpha : [0, +\infty] \to S_1 : \alpha(\cdot)$ is measurable$\}$ denote the set of controls. We denote

by f the speed function, and assume the following basic regularity assumption:

Assumption 2.1.

i. $f : \overline{\Omega} \times S_1 \to (0, +\infty)$ is continuous.
ii. There exists positive constants $L_f, \gamma, L_{f,\alpha}$ such that $|f(x,\alpha) - f(x',\alpha)| \leqslant L_f|x - x'|$, and $|f(x,\alpha) - f(x,\alpha')| \leqslant L_{f,\alpha}|\alpha - \alpha'|^\gamma, \forall x, x' \in \Omega, \alpha, \alpha' \in S_1$.

Our goal is to find the minimum time necessary to travel from the source point $x_{\rm src} \in \Omega$ to the destination point $x_{\rm dst} \in \Omega$, and the optimal trajectories, together with the optimal control α. We consider the controlled dynamical system:

$$\dot{x}(t) = f(x(t), \alpha(t))\alpha(t) \ . \tag{1}$$

We denote by $x_{\alpha,x_{\rm src}}$, or simply x_α, the solution of (1) with $\alpha \in \mathcal{A}$, such that $x_\alpha(0) = x_{\rm src}$ and $x_\alpha(s) \in \overline{\Omega}$ for all $0 \leqslant s \leqslant t$. We restrict the set of control trajectories so that the state x_α stays inside the domain $\overline{\Omega}$ forever, i.e., we consider the following set of admissible controls trajectories:

$$\mathcal{A}_{\Omega,x_{\rm src}} := \{\alpha \in \mathcal{A} \mid x_{\alpha,x_{\rm src}}(s) \in \overline{\Omega}, \forall s \geqslant 0\},$$

and we further assume $A_{\Omega,x_{\rm src}} \neq \emptyset$. In other words, the structure of $A_{\Omega,x_{\rm src}}$ is adapted to the state constraint $x_\alpha(s) \in \overline{\Omega}$. By doing so, the minimum time function can be defined as

$$T_{\rm s\to}(x) = \inf_{\alpha \in \mathcal{A}_{\Omega,x_{\rm src}}} \inf\{\tau \mid x_{\alpha,x_{\rm src}}(\tau) = x\} \ .$$

Consider the Kruzkov change of variable:

$$v_{\rm s\to}(x) = 1 - e^{-T_{\rm s\to}(x)} \ .$$

Then, $v_{\rm s\to}(x)$ is the viscosity solution of the following state constrained HJB equation:

$$\begin{cases} F(x, v_{\rm s\to}(x), Dv_{\rm s\to}(x)) = 0, \ x \in \Omega, \\ F(x, v_{\rm s\to}(x), Dv_{\rm s\to}(x)) \geqslant 0, \ x \in \partial\Omega, \\ v_{\rm s\to}(x_{\rm src}) = 0 \ . \end{cases} \tag{2}$$

where $F(x, r, p) = -\min_{\alpha \in S_1}\{-p \cdot f(x,\alpha)\alpha + 1 - r\}$.

3. THE OPTIMAL TRAJECTORIES OF THE CONTINUOUS SPACE PROBLEM

In this section we show how to reduce the state space Ω of the original minimum time problem, while preserving the optimal trajectories.

Definition 3.1. For every $x, y \in \Omega$, we denote by $\Gamma(x,y)$ the set of *geodesic points* from x to y, defined as

$$\Gamma(x,y) = \{x_{\alpha,x}(s) \mid s \in [0,\tau], \alpha \in \mathcal{A}_{\Omega,x}, x_{\alpha,x}(\tau) = y \text{ and } \int_0^\tau e^{-s}ds = \inf_{\tau \geqslant 0, \alpha \in \mathcal{A}_{\Omega,x},\ x_{\alpha,x}(\tau)=y}\{\int_0^\tau e^{-s}ds\}\}.$$

In other words, z is a geodesic point between x and y if an optimal trajectory from x to y passes through z.

Let us consider a new minimal time optimal control problem ending in $x_{\rm dst}$ with the same dynamics as (1), but starting at any $x \in \overline{\Omega}$. The associated minimal time function is

$$T_{\to\rm t}(x) = \inf_{\alpha \in \mathcal{A}_{\Omega,x}} \inf\{\tau \mid x_{\alpha,x}(\tau) = x_{\rm dst}\} \ .$$

By doing so, we have $T_{\rm s\to}(x_{\rm dst}) = T_{\to\rm t}(x_{\rm src})$, and we denote $\tau^* = T_{\rm s\to}(x_{\rm dst}) = T_{\to\rm t}(x_{\rm src})$. We then use the same change of variable technique to get $v_{\to\rm t}(x) = 1 - e^{-T_{\to\rm t}(x)}$.

Let us denote:

$$\mathcal{F}_v(x) := v_{\rm s\to}(x) + v_{\to\rm t}(x) - v_{\rm s\to}(x)v_{\to\rm t}(x) \ .$$

Let $\Gamma^* = \Gamma(x_{\rm src}, x_{\rm dst})$ denote the union of all optimal trajectories from $x_{\rm src}$ to $x_{\rm dst}$. Then, by the dynamic programming principle, the following result holds :

Lemma 3.1. We have

$$v_{\rm s\to}(x_{\rm dst}) = v_{\to\rm t}(x_{\rm src}) = \inf_{y \in \Omega} \mathcal{F}_v(y) \ . \tag{3}$$

Moreover, if Γ^* is not empty, then for every $x \in \Gamma^*$ we have $\mathcal{F}_v(x) = v_{\rm s\to}(x_{\rm dst})$, that is x is optimal in (3). If there exists an optimal trajectory between any two points of Ω, then x is optimal in (3), that is $\mathcal{F}_v(x) = v_{\rm s\to}(x_{\rm dst})$ if and only if $x \in \Gamma^*$.

Let us now consider an open subdomain of Ω, $\mathcal{O}_\eta \subseteq \Omega$, determined by a parameter $\eta > 0$, and defined as follows:

$$\mathcal{O}_\eta = \{x \in \Omega \mid \mathcal{F}_v(x) < \inf_{y \in \Omega}\{\mathcal{F}_v(y) + \eta\}\} \ . \tag{4}$$

We intend to reduce the state space of our original optimal control problem from Ω to $\mathcal{O}_\eta$. More precisely, we consider a new optimal control problem with the same dynamics, but we restrict the controls so that the state x_α stays inside the domain $\mathcal{O}_\eta$, leading to the new set of controls:

$$\mathcal{A}_{\eta,x_{\rm src}} := \{\alpha \in \mathcal{A} \mid x_{\alpha,x_{\rm src}}(s) \in \overline{O_\eta}, \forall s \geqslant 0\}.$$

Let $v^\eta_{\rm s\to}(x)$ denote the value function of the new problem, then $v^\eta_{\rm s\to}(x)$ is a viscosity solution of the following HJB equation:

$$\begin{cases} F(x, v^\eta_{\rm s\to}(x), Dv^\eta_{\rm s\to}(x)) = 0, \ x \in O_\eta, \\ F(x, v^\eta_{\rm s\to}(x), Dv^\eta_{\rm s\to}(x)) \geqslant 0, \ x \in \partial O_\eta, \\ v^\eta_{\rm s\to}(x_{\rm src}) = 0 \ . \end{cases} \tag{5}$$

Then we have the following result:

Proposition 3.1. If Γ^* is not empty, then $\Gamma^* \subseteq \mathcal{O}_\eta$, and for all $x \in \Gamma^*$ we have $v_{\rm s\to}(x) = v^\eta_{\rm s\to}(x), v_{\to\rm t}(x) = v^\eta_{\to\rm t}(x)$.

The above results express properties of exact optimal trajectories. We will also consider approximate, $\delta-$optimal, trajectories:

Definition 3.2. For every $x, y \in \Omega$, for any $\delta > 0$, we denote $\Gamma_\delta(x,y)$ the set of $\delta-$geodesic points from x to y , defined as :

$$\Gamma_\delta(x,y) = \{x_{\alpha,x}(s) \mid s \in [0,\tau], \alpha \in \mathcal{A}_{\Omega,x}, x_{\alpha,x}(\tau) = y, \text{ and } \int_0^\tau e^{-s}ds \leqslant \inf_{\tau \geqslant 0, \alpha \in \mathcal{A}_{\Omega,x},\ x_{\alpha,x}(\tau)=y}\{\int_0^\tau e^{-s}ds\} + \delta\} \ .$$

Let $\Gamma^*_\delta = \Gamma_\delta(x_{\rm src}, x_{\rm dst})$ denote the set of all $\delta-$geodesic points from $x_{\rm src}$ to $x_{\rm dst}$. Then we have the following results:

Lemma 3.2. For every $\eta > \delta > 0$, we have $\Gamma^*_\delta \subseteq \mathcal{O}_\eta$.

Lemma 3.3. For every $\eta > 0$, there exists $\delta' < \eta$ such that $\mathcal{O}_\eta \subseteq \Gamma^*_{\eta+\delta'}$.

The above two lemmas entail that the sets of $\delta-$geodesic points and $\mathcal{O}_\eta$ constitute two equivalent families of neighborhoods. Thanks to these properties, we establish the following result :

Theorem 3.1. For every $x \in \Gamma^*_\delta$, we have

$$v^\eta_{\rm s\to}(x) = v_{\rm s\to}(x), \quad v^\eta_{\to\rm t}(x) = v_{\to\rm t}(x) \ .$$

Thus, if we are only interested to find $v_{s\to}(x_{\text{dst}})$ and the optimal trajectories between x_{src} and x_{dst}, we only need to solve the reduced problem (5) in the subdomain $\mathcal{O}_\eta$. This is equivalent to solving the original problem in Ω as long as $\mathcal{O}_\eta$ contains the optimal trajectories of this problem.

4. THE MULTI-LEVEL FAST-MARCHING ALGORITHM

4.1 Motivation

The main idea of our algorithm is based on the property proposed above. Instead of computing the optimal trajectories directly, we first find an approximate subdomain of O_η by applying the fast marching search in a coarse grid, with relaxed accuracy and error requirements. Then, we further discretize O_η using a finer grid, and perform a search on this fine grid.

4.2 The Update Operator for The Fast-Marching Search

Based on Assumption 2.1, there exist constants $\underline{f}, \overline{f}$ such that: $0 < \underline{f} \leqslant f(x,\alpha) \leqslant \overline{f} < \infty$. We define $\Upsilon := \frac{\overline{f}}{\underline{f}} \geqslant 1$, and observe that this constant can be interpreted as a measure of anisotropy of the minimum time problem.

Let $X = \Omega \cap (h\mathbb{Z})^d$, we define $V_{s\to}$ as the approximation of the value function $v_{s\to}$ in X. Let $x \in X$, for any I adjacent nodes $x_1, x_2, ..., x_I$ in the grid X, we define $x_\rho = \sum_{i=1}^I \rho_i x_i$ for $\rho \in \Delta^I = \{\rho_i \geqslant 0, \sum_{i=1}^I \rho_i = 1\}$. Let us denote $d(\rho) = \|x_\rho - x\|$ and $\alpha_\rho = \frac{x_\rho - x}{\|x_\rho - x\|}$, which are the distance and the direction from x to x_ρ.

Let $\mathcal{I} = \{1, 2, ..., I\}$. Let $V_{s\to}(x; (x_i)_{i\in\mathcal{I}})$ denote the approximate value $V_{s\to}(x)$ of $v_{s\to}(x)$ depending on the nodes $(x_i)_{i\in\mathcal{I}}$. Then, applying a semi-Lagrangian discretization scheme, we have:

$$V_{s\to}(x; (x_i)_{i\in\mathcal{I}}) = \min_{\rho\in\Delta^I}\{(1 - \frac{d(\rho)}{f(x,\alpha_\rho)}) \sum_i (\rho_i V_{s\to}(x_i)) + \frac{d(\rho)}{f(x,\alpha_\rho)}\} .$$

Let us denote $N(x)$ the set of neighborhood nodes of x, defined as follows:

$$N(x) := \{x_j \in X \mid \exists x_k \in X, s.t. \ \exists \tilde{x} \in [x_j, x_k],\ \|\tilde{x} - x\| \leqslant \Upsilon h\},$$

where we use $[a, b]$ to denote the line segment between a and b. Then, the update operator for the value function in the fast-marching method is as follows:

$$\mathcal{U}(V_{s\to}(x)) := \min\{V_{s\to}(x), \min_{(x_i)_{i\in\mathcal{I}}\in N(x)} V_{s\to}(x; (x_i)_{i\in\mathcal{I}})\} .$$

4.3 The Algorithm

Two Level Fast Marching. We start by describing the special case of the algorithm with only two levels of grid. The algorithm consists of three main steps: Step 1. Discretize Ω using the grid X^H, and find a good $O_\eta^{H,I} \subseteq \Omega$. Step 2. Discretize $O_\eta^{H,I}$ using the fine grid O_η^h. Step 3. Doing a fast marching search in grid O_η^h. We give the details of the first two steps:

Step 1, discretization in coarse grid. We start with the full domain Ω. Let $X^H = \Omega \cap (H\mathbb{Z})^d$. Without loss of generality, we assume $x_{\text{src}}, x_{\text{dst}} \in X^H$.

By doing a fast marching search in X^H starting from x_{src} and x_{dst} respectively, we get the numerical approximation $V_{s\to}^H$ for $v_{s\to}$, and $V_{\to t}^H$ for $v_{\to t}$. We use the approximate value functions $V_{s\to}^H$ and $V_{\to t}^H$ to construct a subset of X^H, denoted by O_η^H:

$$O_\eta^H = \{x^H \in X^H \mid \mathcal{F}_{V^H}(x^H) \leqslant \min_{x^H\in X^H} \mathcal{F}_{V^H}(x^H) + \eta^H\}.$$

Then we construct a continuous analogue of the grid O_η^H, denoted by $O_\eta^{H,I}$, by defining the approximate value functions $V_{s\to}$ and $V_{\to t}$ on the whole domain using linear interpolation of the functions $V_{s\to}^H$ and $V_{\to t}^H$.

Step 2, discretization in fine grid-h. Let the finite set X^h denote the approximation of Ω with mesh step h. Without loss of generality we assume $x_{\text{src}}, x_{\text{dst}} \in X^h$. Let the finite set O_η^h be the approximation of $O_\eta^{H,I}$ with the mesh step h, given by:

$$O_\eta^h = \{x^h \in X^h \mid \exists x^H \in O_\eta^H : \|x^h - x^H\| \leqslant \max((H-h), h)\}.$$

Multi-Level Grids. The computation in the 2-level method above can be easily extended to the Multi-Level case. For each level l, we consider the optimal control problem in which the domain Ω is restricted to the fine-grid region $O_{\eta^{l-1}}^{H_{l-1},I}$ of the previous level $l-1$. Then, by applying the first two steps of the 2-level algorithm as above, we get a new domain, $O_{\eta^l}^{H_l,I}$, taken to be the new state space for the optimal control problem to be solved at the next level. At the end, we do a fast-marching search starting from x_{src} in the final fine grid.

4.4 Implementation

Algorithm 1 Two-Level Fast-Marching Method

Input: Mesh step of coarse and fine grids: H, h. The parameter η^H. The update operator for fast-marching method: $\mathcal{U}$. Start and end point: $x_{\text{src}}, x_{\text{dst}}$.
Output: Approximated value function: $V_{s\to}^h(x)$.

```
Do the fast-marching search staring from x_src and x_dst.
for Every node x^H ∈ X^H do
    if F_{V^H}(x) ⩽ min_{x^H∈X^H} F_{V^H}(x^H) + η^H then
        Set x^H as ACTIVE , store it's position.
    end if
end for
Begin with set FINE be empty.
for Every node x^H in the ACTIVE set do
    for Every x^h ∈ X^h : ‖x^h − x^H‖_∞ ⩽ max{H − h, h} do
        if x^h does not exist in set FINE then
            Add x^h in the set FINE, store it's position.
        end if
    end for
end for
Do fast-marching search starting from x_src in FINE.
```

In Algorithm 2, the selection of active nodes corresponds to Step 1 in the two-level algorithm, the selection of fine grid corresponds to Step 2 in the two-level algorithm.

Algorithm 2 Multi-Level Fast-Marching Method

Input: The parameter H_l, η_l, for $l \in \{1, 2, ..., N\}$. The update operator for fast-marching scheme: $\mathcal{U}$. Start and end point: $x_{\rm src}, x_{\rm dst}$.

Output: Approximated value function: $V^h_{\rm s\to}(x)$.

Let X^{H_1} be the COARSE-GRID.
for $l = 1$ to $N - 1$ **do**
    Do the partial fast marching search starting from $x_{\rm src}$ and $x_{\rm dst}$.
    Select the ACTIVE nodes from the COARSE-GRID.
    Select the set FINE nodes based on the ACTIVE nodes, as in Algorithm 1.
    Let the FINE be the new COARSE-GRID.
end for
Do the fast-marching search starting from $x_{\rm src}$ in FINE.

To implement efficiently these algorithms, we need to store the successive constrained grids $O^{H_l}_{\eta_{l-1}}$ in an effective way. In particular, we need to determine if a candidate point $x \in X^{H_l}$ is in the constrained grid $O^{H_l}_{\eta_{l-1}}$, and access to any of its neighbors, in time $O(1)$, while keeping the storage to be in the order of the size of the (constrained) grids. Moreover, we need to store only the points of X^{H_l} that are in the set $O^{H_l}_{\eta_{l-1}}$. These sets are much smaller compared to the sets X^{H_l}, so we maintain them as a collection of nodes (hash table) accessed through a hash function, with a linked list to handle collisions. Each node of the constrained grid $O^{H_l}_{\eta_{l-1}}$ is instrumented with informations providing the value at this node, its position in the grid, and additional variables needed to maintain the set of active nodes in the fast marching propagation. For points of X^{H_l} that are not in the constrained grid $O^{H_l}_{\eta_{l-1}}$, the hash function returns an empty pointer.

5. COMPUTATIONAL COMPLEXITY

In order to choose the parameters of our algorithm, we shall assume that the following error estimation for the approximate value function $V^h_{\rm s\to}$ holds for some $\theta > 0$:

$$|V^h_{\rm s\to}(x) - v_{\rm s\to}(x)| \leqslant Ch^\theta \ . \tag{6}$$

Indeed, by adapting the results of Capuzzo Dolcetta and Ishii (1984), using Assumption 2.1, one can obtain (6) with $\theta = 1/2$. If we further assume that f is of class $\mathcal{C}^2$, then we obtain (6) with $\theta = 1$.

So, our multi-level algorithm need to compute a numerical approximation with an error bound of same order in h, i.e., $\varepsilon \sim Ch^\theta$, for each level. In the two-level case, the two following parameters should be chosen:

1. The mesh step of the coarse grid H.
2. The parameter η^H, which should be big enough for O^H_η to contain the true optimal trajectories.

The estimation (6) give us a theoretical upperbound for η^H. Then, we have the following result:

Proposition 5.1. Assume there exists a finite number of optimal trajectories, and that the distance between $\mathcal{O}_\eta$ and Γ^* is in the order of η^β, with $0 < \beta \leqslant 1$. Then the total computational complexity of the two level fast marching algorithm is

$$\mathcal{C}(H, h) \sim \kappa((\frac{D}{H})^d + (\frac{(\eta^H)^\beta}{h})^{d-1}\frac{D}{h})$$

where D denotes the Euclidean distance from $x_{\rm src}$ to $x_{\rm dst}$.

Note that the "stiffness" parameter β can take any value in $(0, 1]$, as shown in further work. Regarding the error estimation we want to have in fine grid, the parameter H has an optimal value, which we can compute easily.

Similarly, optimizing the parameters of the multi-level fast marching method, and using the error estimation for $V^h_{\rm s\to}$, we get the following result:

Theorem 5.1. With same assumption as Proposition 5.1.

(i) In order to have an error bound ε, we shall take $h = C\varepsilon^2, H_l = C(H_{l+1})^{\frac{2^{l+1}-2}{2^{l+1}-1}}, \forall l \in \{2, 3, ..., N-1\}$. In this case the total computational complexity of our N-level algorithm is in the order of $(\frac{1}{\varepsilon})^{\frac{2}{1+\beta}d}$.

(ii) Suppose now $f \in \mathcal{C}^2$ or $\theta = 1$ in (6), and $\beta < 1$. Then, we shall take $h = C\varepsilon$ and $H_1 = Ch^{\frac{1}{N}}$, $H_l = C(H_{l+1})^{\frac{l}{l+1}}, \forall l \in \{1, 2, ..., N-1\}$. In this case, the total computational complexity of our N-level algorithm in the order of $(\frac{1}{\varepsilon})^{(1-\beta)d}$.

(iii) When $\beta = 1$, set $N = -\lceil d\log(h)\rceil$ and take $\{H_1, H_2, ..., H_{N-1}\}$ as proposed in (ii). Then, the total computational complexity of our algorithm reduces to $\kappa(-C^d d\log(\frac{1}{\varepsilon}))$.

We implememented the algorithm in C++, and ran it on a single core of a Quad-Core IntelCore I7 at 2.3Ghz, with 16Gb of memory. For an eikonal problem on a cubic domain, with obstacles, in dimension 4, with 5 grid levels, the algorithm ran in 59.2s, giving an error of 3.1%, to be compared with a time of 4241.9s for the classical fast marching method, with the same error. Classical fast marching ran out of time from dimension 5. Our algorithm solved a dimension 7 instance in 1300s, with 6 grid levels, leading to an accuracy of 7%.

6. CONCLUSION

We developed a multilevel fast marching algorithm, allowing one to solve an eikonal PDE by reducing the search space to neighborhoods of optimal trajectories. This yields improved complexity bounds, and solves examples of PDE up to dimension 7, for which it leads to a major speedup by comparison with ordinary grid based methods. We believe our method can be extended to HJB PDE of non-eikonal type. To do so, one needs to replace the fast marching sweeps by Bellman-type updates.

REFERENCES

Capuzzo Dolcetta, I. and Ishii, H. (1984). Approximate solutions of the bellman equation of deterministic control theory. *Applied Mathematics and Optimization*, 11(1), 161–181.

Fleming, W.H. and Soner, H.M. (2006). *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media.

Sanders, P. and Schultes, D. (2012). Engineering highway hierarchies. *Journal of Experimental Algorithmics (JEA)*, 17, 1–1.

Sethian, J.A. (1996). A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4), 1591–1595.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Redundant Coordinates and Generalized Momentum Maps in the Optimal Control of Constrained Mechanical Systems

Simeon Schneider, **Peter Betsch**

Institute of Mechanics, Karlsruhe Institute of Technology, Karlsruhe, Germany (e-mail: simeon.schneider@kit.edu, peter.betsch@kit.edu).

Abstract: The optimal control of mechanical systems satisfies an optimal control version of Noether's theorem. Accordingly, there exist generalized momentum maps on the level of the optimal control problem which are conserved if the system has symmetry. For constrained mechanical systems different approaches to define the necessary optimality conditions are known. These approaches will be compared with respect to their capability to preserve the generalized momentum maps. In addition to that, a discretization approach will be proposed which is capable to preserve the generalized momentum maps.

Keywords: Optimal control, symmetry, constrained mechanical systems, differential-algebraic equations, structure-preservation.

1. INTRODUCTION

In the case of constrained mechanical systems the choice of coordinates plays a crucial role. In particular, the choice of coordinates affects the specific form of the equations of motion. In the case of minimal coordinates the equations of motion take the form of nonlinear ordinary differential equations (ODEs) and numerical methods to solve related optimal control problems are well established. On the other hand, the choice of redundant coordinates facilitates the description of general multibody systems. Due to the presence of holonomic constraints the equations of motion take the form of differential-algebraic equations (DAEs). Numerical methods for optimal control problems with DAEs as state equations have not yet reached the level of maturity when compared to optimal control problems with ODEs as state equations.

Using redundant coordinates in the state equations of the optimal control problem affects the necessary optimality conditions. In addition to that, holonomic constraints in the state equations also place restrictions on the boundary conditions of the optimal control problem. Uncontrolled mechanical systems often have symmetry which leads to the conservation of generalized momentum maps on the level of the optimal control problem (Djukić (1973); van der Schaft (1987); Torres (2002)). An analogous statement can be made in the context of constrained mechanical systems (Betsch and Schneider (2021)).

We show that the generalized momentum maps of the optimal control problem can be preserved under discretization. Moreover, we confirm that the numerical results for the optimal control problem in terms of redundant coordinates converge with the results obtained by using minimal coordinates. In this connection, we investigate the role played by different forms of the optimality conditions commonly used in the optimal control of mechanical systems subject to holonomic constraints.

2. STATE EQUATIONS OF CONSTRAINED MECHANICAL SYSTEMS

The present work deals with the optimal control of mechanical systems subject to holonomic constraints. The motion of such systems is governed by differential-algebraic equations (DAEs). The DAEs are typically in index-3 Hessenberg form. For some applications minimal coordinates can be used along with projection methods to eliminate the holonomic constraints, see e.g. Leyendecker et al. (2010). However, using minimal coordinates may lead to coordinate singularities. To prevent the singularities, coordinate switching can be necessary which, on the other hand, is highly inconvenient in the solution of optimal control boundary value problems. In contrast to that, using redundant coordinates associated with the underlying DAEs may provide a general and singularity-free description of the state equations of constrained mechanical systems.

A simple but representative example of a constrained mechanical system is depicted in Figure 1. The configuration manifold of the mathematical pendulum is just the two-dimensional sphere. Thus, the motion can be either described in minimal coordinates by two angles or in redundant coordinates $\mathbf{q} = (q_1, q_2, q_3)$ along with the constraint

$$g_1(\mathbf{q}) = \frac{1}{2}\left(\mathbf{q}^T\mathbf{q} - l_0^2\right) = 0$$

where l_0 is the length of the pendulum. As already mentioned, by using minimal coordinates there will always be a singular point at the pole of the sphere depending on the choice of the minimal coordinates. These coordinate singularities and the problems arising in the optimal control problem can be circumvented easily by using redundant coordinates instead, which point out the importance of redundant coordinates in optimal control problems.

Using Livens principle, the enhanced action integral reads

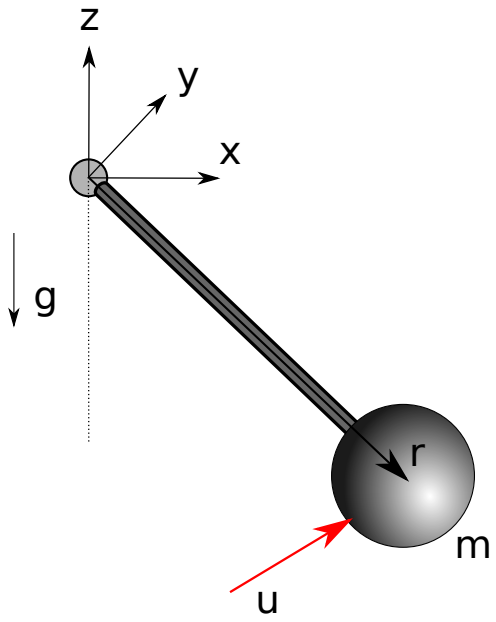


Fig. 1. The controlled mathematical pendulum

$$S = \int_0^T T(\mathbf{q}, \mathbf{v}) - V(\mathbf{q}) - \mathbf{p}^T(\mathbf{v} - \dot{\mathbf{q}}) - \mathbf{y}^T\mathbf{g}(\mathbf{q})\, \mathrm{d}t \quad (1)$$

with $\mathbf{q} \in \mathbb{R}^n$ being the position on the configuration manifold, $\mathbf{v} \in \mathbb{R}^n$ being the velocities at the configuration manifold, $\mathbf{p} \in \mathbb{R}^n$ being the conjugate momenta and $\mathbf{y} \in \mathbb{R}^m$ being the Lagrangian multipliers. Actuating forces can be taken into account by adding their contribution to (1) in the spirit of d'Alembert's principle. Now taking the first variation of (1) yields the equations of motion in the form

$$\left.\begin{aligned} \dot{\mathbf{q}} &= \frac{\partial H(\mathbf{q},\mathbf{p},\mathbf{y})}{\partial \mathbf{p}} \\ \dot{\mathbf{p}} &= -\frac{\partial H(\mathbf{q},\mathbf{p},\mathbf{y})}{\partial \mathbf{q}} + \mathbf{F}(\mathbf{x},\mathbf{u}) \end{aligned}\right\} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},\mathbf{y},\mathbf{u}) \quad (2)$$

$$\mathbf{0} = \mathbf{g}(\mathbf{q}) \quad (3)$$

with the Hamiltonian given by

$$H(\mathbf{q},\mathbf{p},\mathbf{y}) = T(\mathbf{q},\dot{\mathbf{q}}) + V(\mathbf{q}) + \mathbf{y}^T\mathbf{g}(\mathbf{q}) \quad (4)$$

In (2) the state variables are summarized in the compact form $\mathbf{x} = (\mathbf{q}, \mathbf{p}) \in P$, where $P = \mathbb{R}^n \times \mathbb{R}^n$ and $F(\mathbf{x}, \mathbf{u}) : \mathbb{R}^n \times \mathbb{R}^{n_u} \to \mathbb{R}^n$ accounts for the actuating forces.

3. OPTIMAL CONTROL PROBLEM

Now let the optimal control problem seek to minimize the cost function

$$\int_0^T \mathbb{L}(\mathbf{x},\mathbf{u})\, \mathrm{d}t \quad (5)$$

subject to the state equations (2) and (3), which need to be satisfied throughout the time interval $[0, T]$. In (5), $\mathbb{L} : P \times \mathbb{R}^{n_u} \mapsto \mathbb{R}$ is the cost density function. The necessary conditions of optimality may be defined within the Hamiltonian formalism in analogy to Livens principle by using the Pontryagin maximum principle. However, there are different approaches to define the Hamiltonian of the optimal control problem. The simple approach incorporates the constraints (3) directly leading to

$$\tilde{\mathbb{H}}(\mathbf{x},\mathbf{y},\mathbf{u},\boldsymbol{\lambda},\boldsymbol{\eta}) = \boldsymbol{\lambda}^T\mathbf{f}(\mathbf{x},\mathbf{y},\mathbf{u}) + \boldsymbol{\eta}^T\mathbf{g}(\mathbf{q}) - \mathbb{L}(\mathbf{x},\mathbf{u}) \quad (6)$$

On the other hand, according to Roubíček and Valášek (2002) the Hamiltonian for the optimal control problem reads

$$\tilde{\mathbb{H}}(\mathbf{x},\mathbf{y},\mathbf{u},\boldsymbol{\lambda},\boldsymbol{\eta}) = \boldsymbol{\lambda}^T\mathbf{f}(\mathbf{x},\mathbf{y},\mathbf{u}) + \boldsymbol{\eta}^T\mathbf{G}(\mathbf{x},\mathbf{y},\mathbf{u}) - \mathbb{L}(\mathbf{x},\mathbf{u}) \quad (7)$$

where $\mathbf{G}(\mathbf{x},\mathbf{y},\mathbf{u}) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}(\mathbf{g}(\mathbf{q}))$. Supposing the Hamiltonian and the cost function have symmetry, then there exists an optimal control version of Noether's theorem. This holds true for state equations both in the form of ordinary differential equations (Betsch and Becker (2017)) and DAEs (Betsch and Schneider (2021)). Accordingly, if the optimal control problem has symmetry, an associated generalized momentum map is conserved along the solution of the optimal control problem.

In the talk the different approaches will be compared with respect to the conservation of the generalized momentum map and a conserving discretization method will be proposed.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 442997215. This support is gratefully acknowledged.

REFERENCES

Betsch, P. and Becker, C. (2017). Conservation of generalized momentum maps in mechanical optimal control problems with symmetry. *Int. J. Numer. Meth. Engng*, 111(2), 144–175.

Betsch, P. and Schneider, S. (2021). Conservation of generalized momentum maps in the optimal control of constrained mechanical systems. *IFAC PapersOnLine*, 54(9), 615–619.

Djukić, D. (1973). Noether's theorem for optimum control systems. *Int. J. Control*, 18(3), 667–672.

Leyendecker, S., Ober-Blöbaum, S., Marsden, J., and Ortiz, M. (2010). Discrete mechanics and optimal control for constrained systems. *Optim. Control Appl. Meth.*, 31(6), 505–528.

Roubíček, T. and Valášek, M. (2002). Optimal control of causal differential–algebraic systems. *J. Math. Anal. Appl.*, 269(2), 616–641.

Torres, D. (2002). On the Noether theorem for optimal control. *European Journal of Control*, 8(1), 56–63.

van der Schaft, A. (1987). Symmetries in optimal control. *SIAM J. Control and Optimization*, 25(2), 245–259.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Port-Hamiltonian system nodes

Friedrich Philipp *, **Timo Reis** * and **Manuel Schaller** *

* *Institute for Mathematics, Faculty of Mathematics and Natural Sciences, Technische Universität Ilmenau, Ilmenau, Germany (e-mail: {friedrich.philipp,timo.reis,manuel.schaller}@tu-ilmenau.de)*

Abstract: We present a framework to formulate infinite dimensional port-Hamiltonian systems by means of system nodes, which provide a very general and powerful setting for unbounded input and output operators that appear, e.g., in the context of boundary control or observation. One novelty of our approach is that we allow for unbounded and not necessarily coercive Hamiltonian energies. To this end, we construct finite energy spaces to define the port-Hamiltonian dynamics and give an application in case of multiplication operator Hamiltonians where the Hamiltonian density does not need to be positive or bounded. In order to model systems involving differential operators on these finite energy spaces, we show that if the total mass w.r.t. the Hamiltonian density (and its inverse) is finite, one can define a unique weak derivative.

Keywords: Port-Hamiltonian Systems, Infinite Dimensional Systems Theory, Dissipativity

1. INTRODUCTION

The class of port-Hamiltonian (pH) systems provides a flexible and energy-based modeling framework for a wide range of physical systems. Due to their inherent passivity, pH systems are accessible for various control approaches, such as passivity-based control or control by interconnection (van der Schaft and Jeltsema, 2014; van der Schaft, 2000).

Already in finite dimensions there are manifold formulations of pH systems via, e.g. explicit state-space systems (Beattie et al., 2018) or implicit geometric structures (van der Schaft and Jeltsema, 2014). Some formulations could be shown to be equivalent by means of linear relations (Gernandt et al., 2021). In infinite dimensions, classical definitions of port-Hamiltonian systems range from, e.g., the operator theoretic approach of the textbook (Jacob and Zwart, 2012) for distributed parameter systems on one-dimensional domain and its generalization to multidimensional domains (Skrepek, 2021), to implicit definitions with (Stokes)-Dirac structures (van der Schaft and Maschke, 2021, 2002; Schöberl and Siuka, 2014; Le Gorrec et al., 2005; Reis, 2021).

We present a formulation by means of system nodes. This framework poses a generalization of the class of well-posed systems in the sense of Staffans (2005). One feature of system nodes, that is particularly appealing for modeling, is that the generator A and the input map B are not considered separately but rather as a composite operator $A\&B$, allowing for a natural inclusion of the boundary control directly in $\operatorname{dom}(A\&B)$. For results in view of impedance and scattering passivity of system nodes, we refer to Staffans (2002).

In (Villegas, 2007, Section 2.6, Section 3.2), the concepts and relations of boundary control systems and system nodes for scattering or impendance conserving systems with particular focus on port-Hamiltonian systems on one-dimensional domains was analyzed. More precisely, in (Villegas, 2007, Theorem 2.37) the author showed a one-to-one correspondence between port-Hamiltonian boundary control systems with skew-symmetric main operator and system nodes. Further, a relation between Dirac structure and the graph of the system node was provided in (Villegas, 2007, Theorem 3.12).

Here, we also choose the system node approach to port-Hamiltonian systems. We consider multidimensional domains and possible dissipation by means of a dissipative (not conservative) main operator. To this end, we introduce the notion of dissipative system node. Further, we allow for unbounded Hamiltonians or Hamiltonians with non-trivial kernel, which leads to the definition of finite-energy spaces, being the domain of the Hamiltonian function.

Notation. In the following, $(\mathcal{X},\mathcal{U},\mathcal{Y})$ denotes a triple of Hilbert spaces and by $P_{\mathcal{X}}$, $P_{\mathcal{U}}$ and $P_{\mathcal{Y}}$ we denote the orthogonal projection onto $\mathcal{X}$, $\mathcal{U}$ and $\mathcal{Y}$, respectively. By $\operatorname{dom}(S)$ we mean the domain and by $\rho(S) := \{\lambda \in \mathbb{C} \mid \lambda I - S : \operatorname{dom}(S) \to \mathcal{X} \text{ is bijective}\}$ the resolvent set of a possibly unbounded operator S on $\mathcal{X}$. We denote the closed right-half plane by $\mathbb{C}^+ = \{\lambda \in \mathbb{C} \mid \operatorname{Re}\lambda \geq 0\}$ and nonnegative real numbers by $\mathbb{R}^+$.

2. DISSIPATIVE SYSTEM NODES

Consider an unbounded operator $S : \mathcal{X} \times \mathcal{U} \to \mathcal{X} \times \mathcal{Y}$ describing the (abstract) dynamics

$$\begin{bmatrix} \dot{x}(t) \\ -y(t) \end{bmatrix} = S \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}.$$

The concept of operator nodes poses natural assumptions on the operator S, in order to guarantee favorable properties and a suitable solution concept to the abstract

dynamics above. We provide the standard definition given, e.g., in (Opmeer and Staffans, 2014, Definition 2.1).

Definition 1. (Operator Node). An operator node on the triple $(\mathcal{X},\mathcal{U},\mathcal{Y})$ is defined by a (possibly unbounded) linear operator $S : \left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{U}\end{smallmatrix}\right] \to \left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{Y}\end{smallmatrix}\right]$ that is decomposed into $S = \left[\begin{smallmatrix}A\&B\\ C\&D\end{smallmatrix}\right]$ with $A\&B = P_{\mathcal{X}}S : \operatorname{dom}(S) \to \mathcal{X}$ and $C\&D = P_{\mathcal{Y}}S : \operatorname{dom}(S) \to \mathcal{Y}$. Further, we set $Ax = A\&B\left[\begin{smallmatrix}x\\ 0\end{smallmatrix}\right]$ and $\operatorname{dom}(A) = \{x \in \mathcal{X} \mid \left[\begin{smallmatrix}x\\ 0\end{smallmatrix}\right] \in \operatorname{dom}(S)\}$ and demand the following conditions:

(i) $S : \left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{U}\end{smallmatrix}\right] \to \left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{Y}\end{smallmatrix}\right]$ with domain $\operatorname{dom}(S)$ is closed.
(ii) $A\&B : \left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{U}\end{smallmatrix}\right] \to \mathcal{X}$ with domain $\operatorname{dom}(S)$ is closed
(iii) For any $u \in \mathcal{U}$ there is $x \in \mathcal{X}$ with $\left[\begin{smallmatrix}x\\ u\end{smallmatrix}\right] \in \operatorname{dom}(S)$.
(iv) $\operatorname{dom}(A)$ is dense in $\mathcal{X}$ and A has nonempty resolvent set. □

The operator A is also called the *main operator* of the operator node, cf. Opmeer and Staffans (2014).

Definition 2. (System Node). An operator node is called a system node, if its main operator A generates a strongly continuous semigroup on $\mathcal{X}$. □

In the following, we will consider the case $\mathcal{Y} = \mathcal{U}$ as usual for passive systems. Recall that a densely defined linear operator $T : X \supset \operatorname{dom} T \to X$ in the Hilbert space X is called *dissipative* if $\operatorname{Re}\langle Tx, x\rangle \leq 0$ for all $x \in \operatorname{dom} T$. It is called *maximally dissipative* if it is dissipative with no proper dissipative extension. By the Lumer-Phillips theorem, the latter is equivalent to $\operatorname{ran}(\lambda - T) = X$ for some $\lambda > 0$. The same theorem shows that T is maximally dissipative if and only if it generates a contraction semigroup. Moreover, a closed and densely defined operator T in X is maximally dissipative if and only if both T and T^* are dissipative.

Proposition 3. For a linear operator $S : \left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{U}\end{smallmatrix}\right] \to \left[\begin{smallmatrix}\mathcal{X}\\ \mathcal{U}\end{smallmatrix}\right]$ the following statements are equivalent:

(i) S is a dissipative system node.
(ii) S is a maximally dissipative system node.
(iii) S is a maximally dissipative operator node.
(iv) S is a dissipative operator node with main operator A satisfying $\operatorname{ran}(\lambda - A) = \mathcal{X}$ for some $\lambda > 0$.

Proof. (i)⇒(ii). This is a consequence of Lemma 5 and (Staffans, 2002, Lemma 4.3).

(ii)⇒(iii). This is trivial.

(iii)⇒(iv). The operator A is dissipative. Indeed,

$$\begin{aligned}\operatorname{Re}\langle Ax, x\rangle_{\mathcal{X}} &= \operatorname{Re}\left\langle A\&B\begin{bmatrix}x\\0\end{bmatrix}, x\right\rangle_{\mathcal{X}}\\ &= \operatorname{Re}\left\langle S\begin{bmatrix}x\\0\end{bmatrix}, \begin{bmatrix}x\\0\end{bmatrix}\right\rangle_{\mathcal{X}\times\mathcal{U}} \leq 0.\end{aligned} \tag{1}$$

Now, by (Malinen et al., 2006, Proposition 2.4), also S^* is an operator node, whose main operator coincides with A^*. As S is maximally dissipative, also S^* is dissipative and so is A^*. Hence, A is maximally dissipative, which implies $\operatorname{ran}(1 - A) = \mathcal{X}$.

(iv)⇒(i). As in (1) it is seen that A is dissipative. The condition in (iv) on A and the Lumer-Phillips theorem hence imply that A is maximally dissipative and thus generates a contraction semigroup. The operator S is thus a system node. □

We briefly recall a suitable solution concept for dynamics governed by system nodes given, e.g., in (Opmeer and Staffans, 2019, Definition 2.3). Let $S = \left[\begin{smallmatrix}A\&B\\ C\&D\end{smallmatrix}\right]$ be an operator node on $(\mathcal{X},\mathcal{U},\mathcal{Y})$. We call the triple

$$\begin{bmatrix}x\\u\\y\end{bmatrix} \in \begin{bmatrix}C^1(\mathbb{R}^+;\mathcal{X})\\ C(\mathbb{R}^+;\mathcal{U})\\ C(\mathbb{R}^+;\mathcal{Y})\end{bmatrix}$$

a classical trajectory if for all $t \geq 0$

$$\begin{bmatrix}x(t)\\u(t)\end{bmatrix} \in \operatorname{dom}(S) \quad \text{and} \quad \begin{bmatrix}\dot{x}(t)\\-y(t)\end{bmatrix} = \begin{bmatrix}A\&B\\C\&D\end{bmatrix}\begin{bmatrix}x(t)\\u(t)\end{bmatrix}.$$

The following result gives existence of unique solutions with suitable control functions and initial values.

Lemma 4. ((Opmeer and Staffans, 2019, Proposition 2.4)). Let S be a system node on $(\mathcal{X},\mathcal{U},\mathcal{Y})$. Then, for all initial values $x_0 \in \mathcal{X}$ and controls $u \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}^+;\mathcal{U})$ with $\left[\begin{smallmatrix}x_0\\ u(0)\end{smallmatrix}\right] \in \operatorname{dom}(S)$, there is a unique classical trajectory with $x(0) = x_0$. □

The following result shows that dynamics governed by dissipative system nodes are impedance passive, cf. also (Staffans, 2002, Theorem 3.3)

Lemma 5. Let $S = \left[\begin{smallmatrix}A\&B\\ C\&D\end{smallmatrix}\right]$ be a dissipative system node. Then, classical trajectories (x, u, y) satisfy

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\tfrac{1}{2}\|x(t)\|^2 \leq \operatorname{Re}\langle u(t), y(t)\rangle_{\mathcal{U}}.$$

Proof. We provide the short proof for the sake of illustration:

$$\begin{aligned}&\tfrac{\mathrm{d}}{\mathrm{d}t}\tfrac{1}{2}\|x(t)\|^2\\ &= \operatorname{Re}\langle x(t), \dot{x}(t)\rangle_{\mathcal{X}}\\ &= \operatorname{Re}\left\langle \begin{bmatrix}x(t)\\u(t)\end{bmatrix}, S\begin{bmatrix}x(t)\\u(t)\end{bmatrix}\right\rangle_{\mathcal{X}\times\mathcal{U}} - \operatorname{Re}\left\langle u(t), C\&D\begin{bmatrix}x(t)\\u(t)\end{bmatrix}\right\rangle_{\mathcal{U}}\\ &\leq \operatorname{Re}\langle u(t), y(t)\rangle_{\mathcal{U}}.\end{aligned}$$

□

3. PORT-HAMILTONIAN SYSTEM NODES

In this part, we will introduce the notion of port-Hamiltonian system nodes by means of dissipative system nodes. To this end, let a closed and densely defined positive sesquilinear form h on $\mathcal{X}$ be given. We will not assume that h is coercive or bounded, but we require $h(x, x) > 0$ for all non-zero $x \in \operatorname{dom} h$. The form h then induces the quadratic energy Hamiltonian $\mathcal{H}$ via

$$\mathcal{H}(x) := \tfrac{1}{2} \cdot h(x, x), \qquad x \in \operatorname{dom}(\mathcal{H}) = \operatorname{dom}(h).$$

By means of Kato's second representation theorem (Kato, 2013, Chapter 6.2), there exists a unique non-negative self-adjoint operator H in $\mathcal{X}$ with $\operatorname{dom}(H^{1/2}) = \operatorname{dom}(h)$ such that for all $x \in \operatorname{dom}(H)$, $y \in \operatorname{dom}(h)$, we have

$$h(x, y) = \langle Hx, y\rangle.$$

Note that our positivity condition on h is equivalent to $\ker H = \{0\}$.

If h is coercive and bounded, one can define infinite-dimensional linear port-Hamiltonian systems as

$$\begin{bmatrix}\dot{x}\\-y\end{bmatrix} = \begin{bmatrix}A\&B\\C\&D\end{bmatrix}\begin{bmatrix}Hx\\u\end{bmatrix} = \left(\begin{bmatrix}A\&B\\C\&D\end{bmatrix}\begin{bmatrix}H & 0\\0 & I\end{bmatrix}\right)\begin{bmatrix}x\\u\end{bmatrix}, \tag{2}$$

where $S = \left[\begin{smallmatrix}A\&B\\ C\&D\end{smallmatrix}\right]$ is a dissipative system node on $(\mathcal{X},\mathcal{U})$. It is not hard to see that then also $\left[\begin{smallmatrix}A\&B\\ C\&D\end{smallmatrix}\right]\left[\begin{smallmatrix}H & 0\\0 & I\end{smallmatrix}\right]$ is a

system node so that solution theory for (2) is available. Here we consider energies that are neither coercive nor bounded. We then define the space $\mathcal{X}_h$ as the completion of $\operatorname{dom}(h) = \operatorname{dom} H^{1/2}$ with respect to the norm induced by the energy, i.e.,

$$\|\cdot\|_h := \sqrt{\mathcal{H}(x)} = \sqrt{\tfrac{1}{2}h(x,x)} = \tfrac{1}{\sqrt{2}} \cdot \left\|H^{1/2}x\right\|.$$

The space $\mathcal{X}_h$ will be refered to as the *space of finite energy*. We note that in case of coercive and bounded energy (i.e., $\|\cdot\| \sim \|\cdot\|_h$ on $\operatorname{dom}(h)$) we have $\mathcal{X}_h = \mathcal{X}$. If the form h is coercive but unbounded, we have $\mathcal{X}_h = (\operatorname{dom}(h), \|H^{1/2}\cdot\|)$, which is already complete. It can be shown that H extends naturally to an isometric isomorphism

$$\widetilde{H} : \mathcal{X}_h \to \mathcal{X}_h^*$$

which actually coincides with the Riesz map for the Hilbert space $\mathcal{X}_h$.

3.1 Definitions and the solution concept

Using the space of finite energies introduced before, we can now state the definition of a port-Hamiltonian system node in case the Hamiltonian is positive. The case of non-negative energies is subject to current research.

Definition 6. Let h be a positive symmetric sesquilinear form on $\mathcal{X}$. Then a linear *port-Hamiltonian system with respect to the energy form* h has the state space $\mathcal{X}_h$ and takes the form

$$\begin{bmatrix} \dot{x} \\ -y \end{bmatrix} = \begin{bmatrix} A\&B \\ C\&D \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix},$$

where $\left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ is a dissipative system node on $(\mathcal{X}_h, \mathcal{U})$.

In what follows, let $\mathbb{H} : \mathcal{X}_h \times \mathcal{U} \to \mathcal{X}_h^* \times \mathcal{U}$ denote the isometric isomorphism

$$\mathbb{H} := \begin{bmatrix} \widetilde{H} & 0 \\ 0 & I \end{bmatrix}.$$

We say that an operator $M : \mathcal{X}_h^* \times \mathcal{U} \supset \operatorname{dom}(M) \to \mathcal{X}_h \times \mathcal{U}$ is dissipative, if $M\mathbb{H}$ is dissipative as an operator in $\mathcal{X}_h \times \mathcal{U}$ (or, equivalently, if $\mathbb{H}M$ is dissipative as an operator in $\mathcal{X}_h^* \times \mathcal{U}$).

The next proposition can be proved in a straight-forward manner.

Proposition 7. A linear operator S in $\mathcal{X}_h \times \mathcal{U}$ is a dissipative system node on $(\mathcal{X}_h, \mathcal{U})$ if and only if

$$S = M\mathbb{H},$$

where

$$M : \mathcal{X}_h^* \times U \supset \operatorname{dom}(M) \to \mathcal{X}_h \times U$$

is dissipative and closed and, moreover, satisfies the following properties:

(i) $P_{\mathcal{X}_h} M : \mathcal{X}_h^* \times U \supset \operatorname{dom}(M) \to \mathcal{X}_h$ is closed.
(ii) For any $u \in \mathcal{U}$ there is $x^* \in \mathcal{X}_h^*$ such that $\left[\begin{smallmatrix} x^* \\ u \end{smallmatrix}\right] \in \operatorname{dom}(M)$.
(iii) The operator $F : \mathcal{X}_h^* \supset \operatorname{dom} F \to \mathcal{X}_h$ with $\operatorname{dom} F = \{x^* \in \mathcal{X}_h^* : \left[\begin{smallmatrix} x^* \\ 0 \end{smallmatrix}\right] \in \operatorname{dom} M\}$ and $Fx^* = P_{\mathcal{X}_h} M \left[\begin{smallmatrix} x^* \\ 0 \end{smallmatrix}\right]$ is densely defined.
(iv) There is $\lambda > 0$ such that $\lambda\widetilde{H}^{-1} - F : \operatorname{dom}(F) \to \mathcal{X}_h$ is onto.

In what follows, we shall frequently write M as

$$M = \begin{bmatrix} F\&G \\ K\&L \end{bmatrix},$$

where the operators $F\&G$ and $K\&L$ are defined in the obvious way.

3.2 Multiplication operator Hamiltonians

Let $n \in \mathbb{N}$. Here, we treat the particular case $\mathcal{X} = L^2(0,1)^n \cong L^2((0,1), \mathbb{R}^n)$, where the Hamiltonian is the operator of multiplication with a measurable matrix function $m : (0,1) \to \mathbb{R}^{n\times n}$ such that $m(\xi)$ is symmetric positive definite for a.e. $\xi \in (0,1)$. That is, we have

$$\operatorname{dom} H = \{x \in L^2(0,1)^n : mx \in L^2(0,1)^n\}$$

and $(Hx)(\xi) = m(\xi)x(\xi)$ for $x \in \operatorname{dom} H$ and $\xi \in (0,1)$. It is well known that H is a positive self-adjoint operator in $L^2(0,1)^n$ with

$$\operatorname{dom} H^{1/2} = \{x \in L^2(0,1)^n : m^{1/2}x \in L^2(0,1)^n\}$$

and $(H^{1/2}x)(\xi) = m(\xi)^{1/2}x(\xi)$ for $x \in \operatorname{dom} H^{1/2}$ and $\xi \in (0,1)$.

The Hilbert spaces $\mathcal{X}_h$ and $\mathcal{X}_h^*$ can now be represented as

$$\mathcal{X}_h = \{m^{-1/2}x : x \in L^2(0,1)^n\}$$
$$\text{and} \qquad \mathcal{X}_h^* = \{m^{1/2}x : x \in L^2(0,1)^n\}$$

with inner products

$$\langle m^{-1/2}x, m^{-1/2}y\rangle_{\mathcal{X}_h} = \langle x, y\rangle_2$$
$$\text{and} \qquad \langle m^{1/2}x, m^{1/2}y\rangle_{\mathcal{X}_h^*} = \langle x, y\rangle_2,$$

where $\langle\cdot,\cdot\rangle_2$ denotes the inner product of $L^2(0,1)^n$. Note how the extension $\widetilde{H} : \mathcal{X}_h \to \mathcal{X}_h^*$ of H can be likewise represented as multiplication by m.

Our aim is now to define the domain of differential operators as operators from $\mathcal{X}_h^*$ to $\mathcal{X}_h$. To this end, we need to be able to define a weak derivative on a subset of $\mathcal{X}_h^*$ with values in $\mathcal{X}_h$, that is, for $x \in \mathcal{X}_h^* \cap L^2(0,1)^n$, we want that we recover the integration by parts formula in the pivot space $L^2(0,1)^n$, i.e., for all $\varphi \in C_0^\infty(0,1)^n \cap \mathcal{X}_h^*$ with $\varphi' \in \mathcal{X}_h$:

$$\begin{aligned}\langle x', \varphi\rangle_{\mathcal{X}_h \times \mathcal{X}_h^*} = \langle x', \varphi\rangle_{L^2(0,1)^n} &= -\langle x, \varphi'\rangle_{L^2(0,1)^n} \\ &= -\langle \varphi', x\rangle_{\mathcal{X}_h \times \mathcal{X}_h^*}.\end{aligned}$$

In this regard, we obtain the following result guaranteeing a unique weak derivative.

Proposition 8. If $\|m(\cdot)\|, \|m(\cdot)^{-1}\| \in L^1_{\mathrm{loc}}(0,1)$, then

$$C_0^\infty(0,1)^n \subset \{\varphi \in C_0^\infty(0,1)^n \cap \mathcal{X}_h^* : \varphi' \in \mathcal{X}_h\} \subset \mathcal{X}_h^*$$

and $C_0^\infty(0,1)^n$ is dense in $\mathcal{X}_h^*$. In particular, the weak derivative $x' \in \mathcal{X}_h$ of $x \in \mathcal{X}_h^*$ defined via

$$\langle x', \varphi\rangle_{\mathcal{X}_h \times \mathcal{X}_h^*} = -\langle \varphi', x\rangle_{\mathcal{X}_h \times \mathcal{X}_h^*} \tag{3}$$

for all $\varphi \in C_0^\infty(0,1)^n \cap \mathcal{X}_h^*$ with $\varphi' \in \mathcal{X}_h$, is unique

Proof. The proof follows by straightforward computations and application of the fundamental lemma of calculus of variations.

Example 9. As an example, we consider the wave equation with Dirichlet boundary control on a one-dimensional domain $\Omega = (0, \ell)$.

$$\frac{\partial}{\partial t}\begin{pmatrix}\mathbf{q}(t,\xi)\\ \mathbf{p}(t,\xi)\end{pmatrix} = \left(\begin{bmatrix}0 & \frac{\partial}{\partial\xi}\\ \frac{\partial}{\partial\xi} & 0\end{bmatrix} - \begin{bmatrix}0 & 0\\ 0 & d(\xi)\end{bmatrix}\right)\begin{pmatrix}T(\xi)\mathbf{q}(t,\xi)\\ \frac{1}{\rho(\xi)}\mathbf{p}(t,\xi)\end{pmatrix}$$
$$\begin{pmatrix}\mathbf{q}(0,\xi)\\ \mathbf{p}(0,\xi)\end{pmatrix} = \begin{pmatrix}\mathbf{q}_0(\xi)\\ \mathbf{p}_0(\xi)\end{pmatrix}$$
$$0 = \mathbf{p}(t,0), \quad u(t) = \mathbf{p}(t,\ell),$$
$$y(t) = \mathbf{q}(t,\ell).$$

As state space, we choose $\mathcal{X} = L^2(0,1)^2$, $x = (\mathbf{q},\mathbf{p})$. The Hamiltonian is given by

$$\mathcal{H}(\mathbf{q},\mathbf{p}) = \tfrac{1}{2}\left(\|\sqrt{T(\cdot)}\mathbf{q}\|^2_{L^2(\Omega)} + \|\tfrac{1}{\sqrt{\rho(\cdot)}}\mathbf{p}\|^2_{L^2(\Omega)}\right).$$

Set $Hx = \left[\begin{smallmatrix}T & 0\\ 0 & 1/\rho\end{smallmatrix}\right]x$ for $x \in \mathcal{X}$, $\mathcal{X}_h = (\mathcal{X}, \langle H\cdot,\cdot\rangle)$, and $\widetilde{H} = H$. Further, in order to define spatial derivatives, we asume that the assumptions of Proposition 8 are satisfied, i.e.

$$T, T^{-1}, \rho, \rho^{-1} \in L^1(0,\ell).$$

Note that the condition $\rho \in L^1(0,\ell)$ can be interpreted as a finite total mass assumption.

Further, we set $\mathcal{U} = \mathbb{R}$,

$$F\&G\begin{bmatrix}x\\ u\end{bmatrix} = \left(\begin{bmatrix}0 & \frac{\partial}{\partial\xi}\\ \frac{\partial}{\partial\xi} & 0\end{bmatrix} - \begin{bmatrix}0 & 0\\ 0 & d(\xi)\end{bmatrix}\right)x,$$
$$K\&L\begin{bmatrix}x\\ u\end{bmatrix} = -\mathbf{q}(\ell)$$

and $\operatorname{dom} M = \{((\mathbf{q},\mathbf{p}),u) \in W^{1,2}(\Omega)^2 \times \mathbb{R} : \mathbf{p}(\ell) = u, \mathbf{p}(0) = 0\}$. Dissipativity of $M\mathbb{H}$ can be concluded by straightforward integration by parts arguments. The verification of the remaining assumptions of Proposition 7 are subject to current research, in particular as they have to be checked on the finite energy space.

REFERENCES

Beattie, C., Mehrmann, V., Xu, H., and Zwart, H. (2018). Linear port-Hamiltonian descriptor systems. *Mathematics of Control, Signals, and Systems*, 30(4), 17.

Gernandt, H., Haller, F.E., and Reis, T. (2021). A linear relation approach to port-Hamiltonian differential-algebraic equations. *SIAM Journal on Matrix Analysis and Applications*, 42(2), 1011–1044.

Jacob, B. and Zwart, H.J. (2012). *Linear port-Hamiltonian systems on infinite-dimensional spaces*, volume 223. Springer Science & Business Media.

Kato, T. (2013). *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media.

Le Gorrec, Y., Zwart, H., and Maschke, B. (2005). Dirac structures and boundary control systems associated with skew-symmetric differential operators. *SIAM journal on control and optimization*, 44(5), 1864–1892.

Malinen, J., Staffans, O., and Weiss, G. (2006). When is a linear system conservative? *Quarterly of Applied Mathematics*, 64(1), 61–91.

Opmeer, M.R. and Staffans, O.J. (2014). Optimal control on the doubly infinite continuous time axis and coprime factorizations. *SIAM Journal on Control and Optimization*, 52(3), 1958–2007.

Opmeer, M.R. and Staffans, O.J. (2019). Optimal control on the doubly infinite time axis for well-posed linear systems. *SIAM Journal on Control and Optimization*, 57(3), 1985–2015.

Reis, T. (2021). Some notes on port-Hamiltonian systems on banach spaces. *IFAC-PapersOnLine*, 54(19), 223–229.

Schöberl, M. and Siuka, A. (2014). Jet bundle formulation of infinite-dimensional port-Hamiltonian systems using differential operators. *Automatica*, 50(2), 607–613.

Skrepek, N. (2021). Well-posedness of linear first order port-hamiltonian systems on multidimensional spatial domains. *Evolution Equations & Control Theory*, 10(4), 965–1006.

Staffans, O. (2005). *Well-posed linear systems*, volume 103. Cambridge University Press.

Staffans, O.J. (2002). Passive and conservative continuous-time impedance and scattering systems. part i: Well-posed systems. *Mathematics of Control, Signals and Systems*, 15(4), 291–315.

van der Schaft, A.J. (2000). *L2-gain and passivity techniques in nonlinear control*, volume 2. Springer.

van der Schaft, A.J. and Jeltsema, D. (2014). Port-Hamiltonian systems theory: An introductory overview. *Foundations and Trends in Systems and Control*, 1(2-3), 173–378.

van der Schaft, A.J. and Maschke, B. (2002). Hamiltonian formulation of distributed-parameter systems with boundary energy flow. *Journal of Geometry and physics*, 42(1-2), 166–194.

van der Schaft, A.J. and Maschke, B. (2021). Differential operator Dirac structures. *IFAC-PapersOnLine*, 54(19), 198–203.

Villegas, J.A. (2007). A port-hamiltonian approach to distributed parameter systems. *PhD Thesis, University of Twente*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Essentially Decentralized Interior Point Methods for Optimization in Energy Networks

Alexander Engelmann * **Michael Kaupmann** *
and Timm Faulwasser *

* *Institute for Energy Systems, Energy Efficiency and Energy Economics, TU Dortmund University, Dortmund, Germany (e-mail: {alexander.engelmann, timm.faulwasser}@ieee.org).*

Abstract: This note discusses an essentially decentralized interior point method, which is well suited for optimization problems arising in energy networks. Advantages of the proposed method are guaranteed and fast local convergence for problems with non-convex constraints. Moreover, our method exhibits a small communication footprint and it achieves a comparably high solution accuracy with a limited number of iterations. Furthermore, the local subproblems are of low computational complexity. We illustrate the performance of the proposed method on an optimal power flow problem with 708 buses.

Keywords: Distributed Optimization, Decentralized Optimization, Interior Point Method

1. INTRODUCTION

Distributed and decentralized optimization algorithms are key for the optimal operation of networked systems. [1] Applications range from power systems (Worthmann et al., 2015; Erseghe, 2015), via optimal operation of gas networks (Arnold et al., 2009), to distributed control of data networks (Low and Lapsley, 1999).

Classical distributed optimization algorithms used in the above works are, however, typically guaranteed to converge only for problems with convex constraints. Sufficiently accurate models are often non-linear leading to problems with non-convex constraints. Thus, researchers either apply classical methods without convergence guarantees in a heuristic fashion (Erseghe, 2015), or they rely on simplified convex models (Worthmann et al., 2015). Both approaches come with the risk of computing infeasible solutions, which leads to severe risks in practice. Moreover, the convergence rate of classical distributed algorithms is at most linear (Hong and Luo, 2017; Yang et al., 2019).

Lu and Zhu (2018), Yan et al. (2011), and Engelmann et al. (2019) propose distributed second-order methods with fast—i.e. superlinear—convergence guarantees for non-convex problems. These approaches rely on the exchange of quadratic models of the subproblems, which in turn implies a substantial amount of communication and/or central coordination. Quadratic model exchange can be avoided by a combination of an active-set strategy and techniques from inexact Newton methods (Engelmann et al., 2020). However, the detection of the correct active set is difficult and often numerically unstable.

Decomposition of interior point methods can be achieved by solving Newton steps in a decentralized fashion leading to an overall essentially decentralized algorithm. Interior point methods have the advantage that they avoid an active set detection and simultaneously guarantee fast—i.e. superlinear—local convergence for non-convex problems. This note considers the application of the essentially decentralized interior point method (d-IP) from (Engelmann et al., 2021) to an optimal power flow problem, which arises frequently in power systems.

2. PROBLEM FORMULATION

A common formulation of optimization problems in the context of networked systems is

$$\min_{x_i,\dots,x_{|\mathcal{S}|}} \sum_{i\in\mathcal{S}} f_i(x_i) \tag{1a}$$

$$\text{subject to} \quad g_i(x_i) = 0, \qquad \forall i \in \mathcal{S}, \tag{1b}$$

$$h_i(x_i) \leq 0, \qquad \forall i \in \mathcal{S}, \tag{1c}$$

$$\sum_{i\in\mathcal{S}} A_i x_i = b, \tag{1d}$$

where, $\mathcal{S} = \{1,\dots,|\mathcal{S}|\}$ denotes a set of subsystems, each of which is equipped with an objective function $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ and equality and inequality constraints $g_i, h_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{gi}}, \mathbb{R}^{n_{hi}}$. The matrices $A_i \in \mathbb{R}^{n_c \times n_i}$ and the vector $b \in \mathbb{R}^{n_c}$ are coupling constraints between the subsystems.

[1] We refer to an optimization algorithm as being *distributed* if one has to solve a (preferably cheap) coordination problem in a central entity/coordinator. We denote an optimization algorithm as being *decentralized* in absence of such a coordinator and when the agents rely purely on neighbor-to-neighbor communication (Bertsekas and Tsitsiklis, 1989; Nedić et al., 2018). We call an algorithm *essentially decentralized* if it requires no central coordination but a small amount of central communication. We remark that the definition of distributed and decentralized control differs (Scattolini, 2009).

3. A DECENTRALIZED INTERIOR POINT METHOD

Interior point methods reformulate problem (1) via a logarithmic barrier function and slack variables $v_i \in \mathbb{R}^{n_{hi}}$,

$$\min_{x_1,\dots,x_{|\mathcal{S}|},v_1,\dots,v_{|\mathcal{S}|}} \sum_{i\in\mathcal{S}} f_i(x_i) - \mathbb{1}^\top \delta \ln(v_i) \tag{2a}$$

$$\text{subject to} \quad g_i(x_i) = 0, \qquad \forall i \in \mathcal{S}, \tag{2b}$$

$$h_i(x_i) + v_i = 0, \quad v_i \geq 0, \qquad \forall i \in \mathcal{S}, \tag{2c}$$

$$\sum_{i\in\mathcal{S}} A_i x_i = b. \tag{2d}$$

The variable $\delta \in \mathbb{R}_+$ is a barrier parameter, $\mathbb{1} = (1,\dots,1)^\top \in \mathbb{R}^{n_{hi}}$ and the function $\ln(\cdot)$ is evaluated component-wise. Note that the inequality constraints are replaced by barrier functions. Moreover, (2) and (1) share the same minimizers for $\delta \to 0$.

The main idea of interior point methods is to solve (2) for a decreasing sequence of δ. It is often too expensive to solve (2) to full accuracy—hence one typically performs a hand full Newton steps only (Nocedal and Wright, 2006). In this note, we use a variant which computes only *one* Newton step per iteration.

Next, we give a brief summary of distributed interior point methods; details are given in (Engelmann et al., 2021).

3.1 Decomposing the Newton Step

An exact Newton step $\nabla F^\delta(p)\Delta p = -F^\delta(p)$ applied to the first-order optimality conditions $F^\delta(p) = 0$ of (2) reads

$$\begin{pmatrix} \nabla F_1^\delta & 0 & \dots & \tilde{A}_1^\top \\ 0 & \nabla F_2^\delta & \dots & \tilde{A}_2^\top \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{A}_1 & \tilde{A}_2 & \dots & 0 \end{pmatrix} \begin{pmatrix} \Delta p_1 \\ \Delta p_2 \\ \vdots \\ \Delta\lambda \end{pmatrix} = \begin{pmatrix} -F_1^\delta \\ -F_2^\delta \\ \vdots \\ b - \sum_{i\in\mathcal{S}} A_i x_i \end{pmatrix}, \tag{3}$$

where

$$\nabla F_i^\delta = \begin{pmatrix} \nabla_{xx} L_i & 0 & \nabla g_i(x_i)^\top & \nabla h_i(x_i)^\top \\ 0 & -V_i^{-1} M_i & 0 & I \\ \nabla g_i(x_i) & 0 & 0 & 0 \\ \nabla h_i(x_i) & I & 0 & 0 \end{pmatrix},$$

$M_i = \operatorname{diag}(\mu_i)$, $V_i = \operatorname{diag}(v_i)$, and $\tilde{A}_i = (A_i \ 0 \ 0 \ 0)$, cf. (Nocedal and Wright, 2006, Thm. 12.1). Here, $p = (p_1,\dots,p_{|\mathcal{S}|},\lambda)$ and $p_i = (x_i, v_i, \gamma_i, \mu_i)$, where γ_i, μ_i, and λ are Lagrange multipliers assigned to (2b), (2c), and (2d) respectively. Observe that the optimality conditions $F^\delta(p) = 0$ are parameterized by the barrier parameter δ.

The coefficient matrix in (3) has an arrowhead structure, which we exploit for decomposition. Note that each ∇F_i^δ can be computed based on local information only. Assume that ∇F_i^δ is invertible. Then, one can reduce the KKT system (3) by solving the first S block-rows for Δp_i. Hence,

$$\Delta p_i = -\left(\nabla F_i^\delta\right)^{-1}\left(F_i^\delta + \tilde{A}_i^\top \Delta\lambda\right) \text{ for all } i \in \mathcal{S}. \tag{4}$$

Inserting (4) into the last row of (3) yields

$$\left(\sum_{i\in\mathcal{S}} \tilde{A}_i \left(\nabla F_i^\delta\right)^{-1} \tilde{A}_i^\top\right)\Delta\lambda = \left(\sum_{i\in\mathcal{S}} A_i x_i - \tilde{A}_i \left(\nabla F_i^\delta\right)^{-1} F_i^\delta\right) - b. \tag{5}$$

Define

$$S_i \doteq \tilde{A}_i \left(\nabla F_i^\delta\right)^{-1} \tilde{A}_i^\top, \quad \text{and} \tag{6a}$$

$$s_i \doteq A_i x_i - \tilde{A}_i \left(\nabla F_i^\delta\right)^{-1} F_i^\delta - \frac{1}{|\mathcal{S}|} b. \tag{6b}$$

Then, equation (5) is equivalent to

$$\left(\sum_{i\in\mathcal{S}} S_i\right)\Delta\lambda - \sum_{i\in\mathcal{S}} s_i = S\Delta\lambda - s = 0. \tag{7}$$

Observe that once (7) is solved, one can compute $\Delta p_1,\dots,\Delta p_{|\mathcal{S}|}$ locally in each subsystem based on $\Delta\lambda$ via back-substitution into (4). This way, we are able to solve (3) in a distributed fashion, i.e., we first compute (S_i, s_i) locally and then collect (S_i, s_i) in a coordinator. Solving (7) and distributing $\Delta\lambda$ back to all subsystems $i \in \mathcal{S}$ yields Δp_i by evaluating (4).

3.2 Decentralization

Solving (5) in a central coordinator is typically undesirable due to the large amount of information exchange for large-scale systems and due to safety reasons. Hence, we solve (7) in a decentralized fashion via decentralized inner algorithms.

One can show that S is symmetric and positive-semidefinite. Hence, a decentralized version of the conjugate gradient method (d-CG) is applicable. As an alternative, the use of decentralized inner optimization algorithms is possible by reformulating (7) as a convex optimization problem (Engelmann and Faulwasser, 2021).

Solving (7) to full accuracy by inner algorithms is typically expensive in terms of communication and computation. Thus, we use techniques from inexact Newton methods to terminate inner algorithms early based on the violation of the optimality conditions $F^\delta(p) = 0$, cf. (Nocedal and Wright, 2006, Chap. 7.1) . Doing so, one can save a severe amount of inner iterations—especially in early outer iterations. When $\|F^\delta(p^k)\|$ gets closer to zero, we also force the residual of (7) to become smaller to guarantee convergence to a minimizer.

Updating Stepsize and the Barrier Parameter The barrier parameter δ and the stepsize α for the Newton step $p^{k+1} = p^k + \alpha\Delta p^k$ require a small amount of central communication but no central computation. Indeed, it is possible to compute local surrogates $\{\alpha_i\}_{i\in\mathcal{S}}$ and $\{\delta_i\}_{i\in\mathcal{S}}$ and take their minimal/maximal values over all subsystems to obtain (α, δ).

The Overall Algorithm The overall distributed interior point algorithm is summarized in Algorithm 1. Algorithm 1 has local superlinear convergence guarantees for non-convex problems in case the barrier parameter and the residual in (5) decrease fast enough, cf. (Engelmann et al., 2021, Thm. 2).

4. APPLICATION TO OPTIMAL POWER FLOW

Optimal Power Flow (OPF) problems compute optimal generator set-points in electrical power systems while meeting grid constraints and technical limits (Frank and Rebennack, 2016).

Algorithm 1 Ess. Decentralized Interior Point Method

Initialization: p_i^0 for all $i \in \mathcal{S}$, $\delta^0, \lambda^0, \epsilon$
while $\|F^0(p^k)\|_\infty > \epsilon$ **do**
  compute (S_i^k, s_i^k) locally via (6)
  while residual of (7) too large **do**
    iterate (7) via a decentralized algorithm
  end while
  compute stepsize α^k and update $p^{k+1} = p^k + \alpha \Delta p^k$
  update $\delta^{k+1} < \delta^k$, $k \to k+1$
end while
return $p^\star$

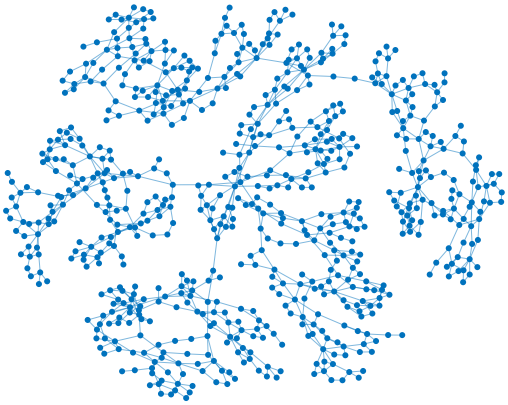


Fig. 1. Six interconnected 118-bus systems.

A basic formulation of the OPF problem reads

$$\min_{s,v \in \mathbb{C}^N} \; f(s) \tag{8a}$$

$$\text{subject to} \quad s - s^d = \operatorname{diag}(v) Y v^*, \tag{8b}$$

$$\underline{p} \le \operatorname{re}(s) \le \bar{p}, \quad \underline{q} \le \operatorname{im}(s) \le \bar{q}, \tag{8c}$$

$$\underline{v} \le \operatorname{abs}(v) \le \bar{v}, \quad v^1 = v^s. \tag{8d}$$

Here, $v \in \mathbb{C}^N$ are complex voltages, and $s \in \mathbb{C}^N$ are complex power injections at all buses N. The operators $\operatorname{re}(\cdot)$ and $\operatorname{im}(\cdot)$ denote the real part and imaginary part of a complex number, and $(\cdot)^*$ denotes the complex conjugate. The objective function $f : \mathbb{C}^N \to \mathbb{R}$ encodes the cost of power generation. The grid physics are described via the power flow equations (8b), where $Y \in \mathbb{C}^{N \times N}$ is the complex bus-admittance matrix describing grid topology and parameters. Moreover, $s^d \in \mathbb{C}^N$ is a fixed power demand. The constraints (8c) describe technical limits on the power injection by generators, and (8d) models voltage limits. The second equation in (8d) is a reference condition on the voltage at the first bus, v^1, where the complex voltage is constrained to a reference value v^s.

Note that one can reformulate the OPF problem (8) in form of (1) by introducing auxiliary variables. Different variants of doing do exist; here we rely on a reformulation from Mühlpfordt et al. (2021).

4.1 A case study

As a case study, we consider 6 interconnected IEEE 118-bus test systems shown in Fig. 1. Each of these systems corresponds to one subsystem $i \in \mathcal{S}$ in problem (1). We use grid parameters from `MATPOWER`, and we interconnect the subsystems in an asymmetric fashion to generate non-zero flows at the interconnection points. In total, we get an optimization problem with about 3.500 decision variables.

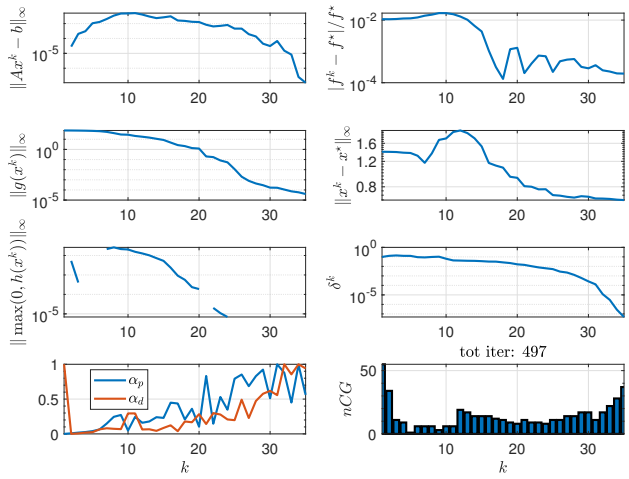


Fig. 2. Convergence of Algorithm 1.

Fig. 2 depicts the convergence of Algorithm 1 over the iteration index k with algorithm parameters from (Engelmann et al., 2021). The figure depicts the consensus violation $\|Ax^k - b\|_\infty$, which can be interpreted as the maximum mismatch of physical values at boundaries between subsystems. Furthermore, the relative error in the objective function $|f^k - f^\star|/f^\star$, the infeasibilities $\|g(x^k)\|_\infty$ and $\|\max(0, h(x^k))\|_\infty$, the distance to the minimizer $\|x^k - x^\star\|_\infty$, the number of inner iterations of d-CG, the barrier parameter sequence $\{\delta^k\}$, and the primal and dual step size (α^p, α^d) are shown. The centralized solution $x^\star$ is computed via the open-source solver `IPOPT` (Wächter and Biegler, 2005).

One can observe that the consensus violation is at the level of 10^{-5} for all iterations. This means that the iterates are feasible with respect to the power transmitted over transmission lines. This results from the fact that the consensus constraint (2d) is implicitly enforced when solving (7) via d-CG. A low consensus violation has the advantage, that one can terminate d-IP early and apply one local NLP iteration to obtain a feasible but possibly suboptimal solution.[2] We note that feasibility is typically of much higher importance than optimality in power systems, since feasibility ensures a safe system operation, cf. Remark 1. From $\|g(x^k)\|_\infty$ and $\|\max(0, h(x^k))\|_\infty$[3] in Fig. 2 one can see that feasibility is ensured to a high degree after 20-30 dIP iterations. At the same time we reach a suboptimality level of almost 0.01%, which is much smaller than in other works on distributed optimization for OPF, cf. (Erseghe, 2015; Guo et al., 2017). Moreover, one can see that the distance to the minimizer $\|x^k - x^\star\|_\infty$ is still quite large due to the small sensitivity of f with respect to the reactive power inputs. This is a well-known phenomenon in the context of OPF problems.

[2] Assuming that the local OPF problem is feasible for the current boundary value iterate.
[3] The blank spots in the plot for $\|\max(0, h(x^k))\|_\infty$ correspond to zero values, since $\log(0) = -\infty$.

Regarding Algorithm 1 itself, one can see that the barrier parameter δ steadily decreases in each iteration. Moreover, during the first 20 iterations, comparably small step-sizes are used. The domain of local convergence is reached after around 30 iterations. Note that we use different stepsizes α_p for the primal variables and α_d for the dual variables. Observe that due to the dynamic termination of inner d-CG iterations based on the inexact Newton theory, Algorithm 1 requires a small amount of inner iterations in the beginning and the number of iterations increases when coming closer to a local minimizer. This effect saves a substantial amount of inner iterations.

The widely used Alternating Direction Method of Multipliers (ADMM) does not converge for the considered case. This seems to occur rarely, but was also reported in other works (Christakou et al., 2017). Algorithm 1 requires 25 seconds for performing 35 iterations with serial execution. The MATPOWER solver `MIPS` needs about 13 seconds when applied to the distributed formulation and 2 seconds when applied to the centralized problem formulation. Executing 497 ADMM iterations—this reflects the number of d-CG iterations in Algorithm 1—requires 210 seconds with serial execution. This illustrates the large computation overhead of ADMM in the local steps, since here one has to solve an NLP in each iteration and for each subsystem. In contrast, d-IP only requires one matrix inversion every outer iteration. All simulations are performed on a standard state-of-the-art notebook.

Remark 1. (Sufficient feasibility in power systems). Note that feasibility in a range of 10^{-3} to 10^{-5} is typically sufficient for a safe power system operation. The parameters in the OPF problem (8), such as power demands and line parameters, induce uncertainty to the problem, which is typically much larger than this level (Kim and Baldick, 2000). Hence, in applications there is typically little-to-no benefit in solving OPF problems to machine precision.

5. SUMMARY & OUTLOOK

We have presented an essentially decentralized interior point method for distributed optimization in energy networks with advantageous properties in terms of convergence guarantees, communication footprint, and practical convergence. We have illustrated the performance of our method on a 708-bus case study. Future work will consider improvements in implementation aspects of d-IP, where we aim faster execution times and at scalability up to several thousand buses.

REFERENCES

Arnold, M., Negenborn, R.R., Andersson, G., and De Schutter, B. (2009). Multi-area predictive control for combined electricity and natural gas systems. In *2009 European Control Conference (ECC)*, 1408–1413. doi:10.23919/ECC.2009.7074603.

Bertsekas, D.P. and Tsitsiklis, J.N. (1989). *Parallel and Distributed Computation: Numerical Methods*, volume 23. Prentice Hall Englewood Cliffs, NJ.

Christakou, K., Tomozei, D.C., Le Boudec, J.Y., and Paolone, M. (2017). AC OPF in radial distribution networks – Part II: An augmented Lagrangian-based OPF algorithm, distributable via primal decomposition. *Electric Power Systems Research*, 150, 24–35. doi:10.1016/j.epsr.2017.04.028.

Engelmann, A. and Faulwasser, T. (2021). Essentially Decentralized Conjugate Gradients. *arXiv: 2102.12311*.

Engelmann, A., Jiang, Y., Houska, B., and Faulwasser, T. (2020). Decomposition of Nonconvex Optimization via Bi-Level Distributed ALADIN. *IEEE Transactions on Control of Network Systems*, 7(4), 1848–1858. doi:10.1109/TCNS.2020.3005079.

Engelmann, A., Jiang, Y., Mühlpfordt, T., Houska, B., and Faulwasser, T. (2019). Toward Distributed OPF Using ALADIN. *IEEE Transactions on Power Systems*, 34(1), 584–594. doi:10.1109/TPWRS.2018.2867682.

Engelmann, A., Stomberg, G., and Faulwasser, T. (2021). An essentially decentralized interior point method for control. In *2021 60th IEEE Conference on Decision and Control (CDC)*, 2414–2420. doi:10.1109/CDC45484.2021.9683694.

Erseghe, T. (2015). A distributed approach to the OPF problem. *EURASIP Journal on Advances in Signal Processing*, 2015(1), 45. doi:10.1186/s13634-015-0226-x.

Frank, S. and Rebennack, S. (2016). An introduction to optimal power flow: Theory, formulation, and examples. *IIE Transactions*, 48(12), 1172–1197. doi:10.1080/0740817X.2016.1189626.

Guo, J., Hug, G., and Tonguz, O.K. (2017). A Case for Nonconvex Distributed Optimization in Large-Scale Power Systems. *IEEE Transactions on Power Systems*, 32(5), 3842–3851. doi:10.1109/TPWRS.2016.2636811.

Hong, M. and Luo, Z. (2017). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2), 165–199.

Kim, B. and Baldick, R. (2000). A comparison of distributed optimal power flow algorithms. *IEEE Transactions on Power Systems*, 15(2), 599–604. doi:10.1109/59.867147.

Low, S. and Lapsley, D. (1999). Optimization flow control. I. Basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7(6), 861–874. doi:10.1109/90.811451.

Lu, Y. and Zhu, M. (2018). Privacy preserving distributed optimization using homomorphic encryption. *Automatica*, 96, 314–325. doi:10.1016/j.automatica.2018.07.005.

Mühlpfordt, T., Dai, X., Engelmann, A., and Hagenmeyer, V. (2021). Distributed power flow and distributed optimization—Formulation, solution, and open source implementation. *Sustainable Energy, Grids and Networks*, 26. doi:10.1016/j.segan.2021.100471.

Nedić, A., Pang, J.S., Scutari, G., and Sun, Y. (2018). *Multi-Agent Optimization: Cetraro, Italy 2014*, volume 2224 of *Lecture Notes in Mathematics*. Springer International Publishing, Cham. doi:10.1007/978-3-319-97142-1.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer Science & Business Media, New York.

Scattolini, R. (2009). Architectures for distributed and hierarchical Model Predictive Control – A review. *Journal of Process Control*, 19(5), 723–731. doi:10.1016/j.jprocont.2009.02.003.

Wächter, A. and Biegler, L.T. (2005). Line Search Filter Methods for Nonlinear Programming: Local Convergence. *SIAM Journal on Optimization*, 16(1), 32–48. doi:10.1137/S1052623403426544.

Worthmann, K., Kellett, C.M., Braun, P., Grüne, L., and Weller, S.R. (2015). Distributed and Decentralized Control of Residential Energy Systems Incorporating Battery Storage. *IEEE Transactions on Smart Grid*, 6(4), 1914–1923. doi:10.1109/TSG.2015.2392081.

Yan, W., Wen, L., Li, W., Chung, C.Y., and Wong, K.P. (2011). Decomposition–coordination interior point method and its application to multi-area optimal reactive power flow. *International Journal of Electrical Power & Energy Systems*, 33(1), 55–60. doi:10.1016/j.ijepes.2010.08.004.

Yang, T., Yi, X., Wu, J., Yuan, Y., Wu, D., Meng, Z., Hong, Y., Wang, H., Lin, Z., and Johansson, K.H. (2019). A survey of distributed optimization. *Annual Reviews in Control*, 47, 278–305. doi:10.1016/j.arcontrol.2019.05.006.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Extremality of Codes in the Lee Metric ⋆

Eimear Byrne * **and Violetta Weger** **

* *School of Mathematics and Statistics,*
University College Dublin Ireland (e-mail: ebyrne@ucd.ie)
** *Department of Electrical and Computer Engineering,*
Technical University of Munich Germany
(e-mail: violetta.weger@tum.de)

Abstract: We investigate Singleton-like bounds in the Lee metric and characterize extremal codes. We then focus on Plotkin-like bounds in the Lee metric and present a new bound that extends and refines a previously known bound, which it out-performs in the case of non-free codes. We then compute the density of codes that meet this bound. Finally, we fill a gap in the characterization of Lee-equidistant codes.

Keywords: Ring-linear code, Lee distance, maximum Lee distance, constant weight code

1. INTRODUCTION

We consider codes over $\mathbb{Z}/p^s\mathbb{Z}$ for the Lee metric. One question of interest concerns the behaviour of maximum Lee distance (MLD) codes, which are those codes that are extremal with respect to Singleton-like bounds in the Lee metric. Two proposals of a Lee-metric Singleton bound are known, namely Shiromoto (2000) and Alderson and Huntemann (2013). We show that independently of which bound one considers, MLD codes are sparse, which is done through a characterization of MLD codes. We also provide an answer to the question of whether dual codes preserve this property, by giving examples of MLD codes whose dual is also MLD as well as counterexamples.

We also consider Plotkin-like bounds in the Lee metric. The first such bound was proposed by Wyner and Graham (1968) and was later improved by Chiang and Wolf (1971). However, the Plotkin bound of Chiang and Wolf holds only for free codes. We thus give a generalization of their bound and in addition obtain an improvement. We then characterize codes attaining this new bound and compute their density.

2. PRELIMINARIES

Throughout this paper we will consider codes to be submodules over the integer residue ring $\mathbb{Z}/p^s\mathbb{Z}$, where p is a prime and s is a positive integer. We write $\langle p^i \rangle$ to denote either the ideal $p^i\mathbb{Z}/p^s\mathbb{Z}$ or the submodule $p^i(\mathbb{Z}/p^s\mathbb{Z})^n$, depending on the context. For the remainder, we write $M := \lfloor p^s/2 \rfloor$. See Honold and Landjev (2000) for a detailed treatment of the topic of linear codes over finite chain rings.

Definition 1. A $\mathbb{Z}/p^s\mathbb{Z}$-module of $(\mathbb{Z}/p^s\mathbb{Z})^n$ is called a linear code of length n.

⋆ The second author is supported by the Swiss National Science Foundation grant number 195290.

In the standard case, over finite fields, a code is a linear subspace of $\mathbb{F}_q^n$ and thus has a dimension k. We take as an analogue of the dimension, the parameter given by:

$$k := \log_{p^s}(|\ \mathcal{C}\ |).$$

It is well known that a $\mathbb{Z}/p^s\mathbb{Z}$ module $\mathcal{C}$ is isomorphic to

$$(\mathbb{Z}/p^s\mathbb{Z})^{k_1} \times \left(\mathbb{Z}/p^{s-1}\mathbb{Z}\right)^{k_2} \times \cdots \times (\mathbb{Z}/p\mathbb{Z})^{k_s}.$$

Therefore, as an additional parameter of the code we call $(k_1, \ldots, k_s)$ its subtype. It holds that

$$k = \sum_{i=1}^{s} \frac{s-i+1}{s} k_i.$$

In addition, k_1 is called the free rank of the code and $K = \sum_{i=1}^{s} k_i$ is called its rank.

Definition 2. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a linear code of rank K. We call any $K \times n$ matrix G a generator matrix of $\mathcal{C}$, if its row-span is $\mathcal{C}$. For the codes considered here, the socle of $\mathcal{C}$ is $\langle p^{s-1} \rangle \cap \mathcal{C}$.

The dual of the linear code $\mathcal{C}$ is denoted by $\mathcal{C}^\perp$ and defined in the usual way, that is:

$$\mathcal{C}^\perp = \{x \in (\mathbb{Z}/p^s\mathbb{Z})^n \mid x \cdot c = 0 \ \text{ for all } c \in \mathcal{C}\}.$$

Proposition 3. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a linear code of order p^{sk}, subtype $(k_1, k_2, \ldots, k_s)$, free rank k_1 and rank K. Then $\mathcal{C}^\perp$ is a linear code of order $p^{s(n-k)}$, subtype $(n - K, k_s, \ldots, k_2)$, free rank $n - K$ and rank $n - k_1$.

In this paper we will focus on the Lee metric. However, this will often be in reference to the Hamming metric. Recall that for $x, y \in (\mathbb{Z}/p^s\mathbb{Z})^n$, the Hamming distance between x and y is defined to be

$$d_H(x, y) = |\{i \in \{1, \ldots, n\} \ \mid \ x_i \neq y_i|\}.$$

Furthermore, the Hamming weight of x is $w_H := d_H(0, x)$. The minimum Hamming distance of $\mathcal{C}$ is $d_H(\mathcal{C}) := \min\{d_H(x, y) \ \mid \ x, y \in \mathcal{C}, x \neq y\}$.

Definition 4. For $x \in \mathbb{Z}/p^s\mathbb{Z}$ we denote by $w_L(x)$ the Lee weight of x, which is defined to be:

$$w_L(x) = \min\{x, p^s - x\},$$

where x is interpreted as an integer in $\{0,\dots,M\}$ in the evalutation $w_L(x)$. For $x \in (\mathbb{Z}/p^s\mathbb{Z})^n$, the Lee weight is defined additively, that is

$$w_L(x) = \sum_{i=1}^{n} w_L(x_i).$$

The Lee weight induces the Lee distance, i.e., for $x, y \in (\mathbb{Z}/p^s\mathbb{Z})^n$ we set

$$d_L(x,y) = w_L(x-y).$$

Note that for $x \in (\mathbb{Z}/p^s\mathbb{Z})^n$, $w_H(x) \leq w_L(x) \leq M w_H(x)$.

Definition 5. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a linear code. Then the minimum Lee distance of $\mathcal{C}$ is defined as

$$d_L(\mathcal{C}) = \min\{w_L(c) \mid c \in \mathcal{C}, c \neq 0\}.$$

One can easily observe that

$$d_H(\mathcal{C}) \leq d_L(\mathcal{C}) \leq M d_H(\mathcal{C}).$$

In order to simplify the proofs used in the following bounds we start by making some initial observations and hence unifying the techniques. For the Hamming metric on $\mathbb{Z}/p^s\mathbb{Z}$ the following Singleton-like bounds are known:

Proposition 6. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a code of order p^{sk}, then

$$d_H(\mathcal{C}) \leq n - k + 1.$$

Of course the above bound holds even if the code is not linear. It is well known (see for example Dougherty (2017); Dougherty and Shiromoto (2000)) that for linear codes one can also formulate a tighter bound:

Proposition 7. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a linear code of rank K, then

$$d_H(\mathcal{C}) \leq n - K + 1.$$

Codes that achieve the bound of Proposition 7 are said to have the property of being maximum distance with respect to the rank and are referred to as (MDR) codes. Note that any linear code that is MDS is also MDR and $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ is an MDR code if and only if its socle $\mathcal{C}'$ is an MDS code over $\mathbb{F}_p$.

Let $\mathcal{C}$ be a $\mathbb{Z}/p^s\mathbb{Z}$-module. We define

$$\overline{w}_L(\mathcal{C}) := \frac{1}{|\mathcal{C}|} \sum_{a \in \mathcal{C}} w_L(a),$$

that is, $\overline{w}_L(\mathcal{C})$ is the average Lee weight of the $\mathbb{Z}/p^s\mathbb{Z}$-module $\mathcal{C}$.

A Plotkin-like bound in the Lee metric is:

$$d_L(\mathcal{C}) \leq \frac{|\mathcal{C}|}{|\mathcal{C}|-1} \overline{w}_L(\mathcal{C}). \tag{1}$$

Clearly, this bound is met if and only if every non-zero codeword of $\mathcal{C}$ has the same Lee weight, i.e., if and only if $\mathcal{C}$ is Lee-weight equidistant.

3. OVERVIEW OF EXISTING BOUNDS

3.1 Singleton-like Bounds

For the most well-known case, i.e., $\mathbb{Z}/4\mathbb{Z}$, the Singleton bound is given through the Gray isometry.

Theorem 8. ($\mathbb{Z}/4\mathbb{Z}$-Singleton Bound). Let $\mathcal{C} \subseteq (\mathbb{Z}/4\mathbb{Z})^n$ of order p^{sk}. Then

$$d_L(\mathcal{C}) \leq 2(n-k)+1.$$

Theorem 9. (Shiromoto (2000)). Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ have order p^{sk}. Then

$$\left\lfloor \frac{d_L(\mathcal{C})-1}{M} \right\rfloor \leq n - \lceil k \rceil.$$

Corollary 10. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a linear code of rank K. Then

$$d_L(\mathcal{C}) \leq M(n-K+1).$$

Corollary 11. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a linear code of rank K. Then

$$\left\lfloor \frac{d_L(\mathcal{C})-1}{M} \right\rfloor \leq n - K.$$

If a linear code $\mathcal{C}$ is such that $d_L(\mathcal{C}) = M(n-k+1)$ which is greater or equal to $M(n-K+1)$, we must have that it attains the bound of Corollary 10 as well and that $\mathcal{C}$ is a free MDS code.

One can also consider the bound provided by Alderson and Huntemann (2013).

Theorem 12. For any code $\mathcal{C}$ in $(\mathbb{Z}/p^s\mathbb{Z})^n$ of order p^{sk}, $k \in \{2,\dots,n\}$ we have that

$$d_L(\mathcal{C}) \leq M(n-k).$$

A first Lee-metric Plotkin-like bound was provided by Wyner and Graham (1968) and then extended in Chiang and Wolf (1971). Equidistant codes attain these bounds.

The following generalizes the result of Chiang and Wolf to any (possibly non-free) code.

Proposition 13. Let $\mathcal{C}$ be a linear code in $(\mathbb{Z}/p^s\mathbb{Z})^n$ of subtype $(k_1,\dots,k_s)$, with $k_1 \geq 1$. Then

$$d_L(\mathcal{C}) \leq \begin{cases} \dfrac{p^s+1}{4}(n-k_1+1) \text{ if } p \text{ is odd,} \\ \dfrac{2^{2s-2}}{2^s-1}(n-k_1+1) \text{ if } p=2. \end{cases}$$

3.2 Characterization

We now characterize codes attaining the Singleton-like bounds in the Lee metric.

Proposition 14. The only linear codes that attain the $\mathbb{Z}/4\mathbb{Z}$-Singleton bound of Theorem 8 are $\mathcal{C} = \langle(2,\dots,2)\rangle$, its dual $\mathcal{C}^\perp$ and the ambient space $(\mathbb{Z}/4\mathbb{Z})^n$ itself.

While the Singleton-like bound from Theorem 9 is sharp, there are very few linear codes that attain this bound. We exclude the trivial case $\mathcal{C} = (\mathbb{Z}/p^s\mathbb{Z})^n$ of minimum Lee distance 1, which always attains the bound.

Theorem 15. The only linear codes $\mathcal{C} \subset (\mathbb{Z}/p^s\mathbb{Z})^n$ of order p^{sk} and rank K that attain the bound of Theorem 9 are

- for p odd: codes equivalent to $\mathcal{C} = \langle(1,2)\rangle \subset (\mathbb{Z}/5\mathbb{Z})^2$ or over any p^s with $k < \lceil k \rceil = K = n < k+1$, i.e., $d_L(\mathcal{C}) = 1$;
- for $p = 2$: $\mathcal{C} = \langle(2^{s-1},\dots,2^{s-1})\rangle$ with $d_L(\mathcal{C}) = 2^{s-1}n$, or such that $k \neq K = \lceil k \rceil \in \{n, n-1\}$ giving $d_L(\mathcal{C}) \leq 2^{s-1}$ and $d_L(\mathcal{C}) = 2^s$ respectively.

In (Alderson and Huntemann, 2013, Lemma 13), it was already observed that for $k > 1 \in \mathbb{N}$, there is no linear code that attains the bound of Theorem 9. We have thus extended their characterization. However, also for the bound of Theorem 12 from Alderson and Huntemann (2013), we have that only very few linear codes are extremal:

Theorem 16. The only linear codes $\mathcal{C} \subset (\mathbb{Z}/p^s\mathbb{Z})^n$ of order p^{sk} and rank K that attain the bound of Theorem 12 are

- for p odd: codes with $p^s = 5, k+1 \leq n \leq k+3$ or free codes with $p^s \in \{7, 9\}, n = k+1$;
- for $p = 2$: free codes with $s = 2, k+1 \leq n \leq k+2$, free codes with $s = 3, n = k+1$ or $k+1 = K \in \{n, n-1\}$.

In particular, this implies that in the case of p odd, we must have $d_L(\mathcal{C}) \in \{2, 3, 4, 6\}$ and if $p = 2$ we must have $d_L(\mathcal{C}) \in \{2, 4, 2^{s-1}, 2^s\}$.

Since an extremal code for the bound of Theorem 9 is such that $\lceil k \rceil = K$ and $d_L(\mathcal{C}) = M(n - \lceil k \rceil) + \alpha$ for some $\alpha \in \{1, \ldots, M\}$, Theorem 15 also includes the bound of Corollary 11.

Corollary 17. The only linear codes $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ that attain the bound of Corollary 11 are:

- for p odd: $K = n$ or a code equivalent to $\mathcal{C} = \langle (1, 2) \rangle \subset (\mathbb{Z}/5\mathbb{Z})^2$;
- for $p = 2$: $\mathcal{C} = \langle (2^{s-1}, \ldots, 2^{s-1}) \rangle$, and $K \in \{n, n-1\}$.

We will call a code maximum Lee distance (MLD) if it is extremal code with respect to any of the considered Singleton-like bounds. One can immediately see that the density of MLD codes is 0 for $p \to \infty$. For any fixed rate $R = k/n$ or for any fixed $p > 7$ we can also see that the density of MLD codes is 0 for $n \to \infty$.

Notice, however, that also with more sophisticated bounds the number of maximum Lee distance codes for $p = 2$ will remain small, due to the fact that the socle $\mathcal{C} \cap \langle 2^{s-1} \rangle$ is a trivial binary MDS code, which cannot be avoided.

The characterizations just shown also answer the question as to whether the dual of MLD codes is also MLD; for the special case of $\mathbb{Z}/4\mathbb{Z}$, we have seen that non-trivial linear codes that achieve the $\mathbb{Z}/4\mathbb{Z}$-Singleton bound, i.e., $d_L(\mathcal{C}) \leq 2(n-k)+1$, are only the codes $\mathcal{C} = \langle (2, \ldots, 2) \rangle$ and its dual. Thus, for this particular bound it is true that the dual of an optimal code attains the bound as well, however, for the other Singleton-like bounds, this is (in general) not true, as the choice of n is very restrictive.

4. NEW BOUNDS

Let $\mathcal{C}$ be a $\mathbb{Z}/p^s\mathbb{Z}$-submodule of $(\mathbb{Z}/p^s\mathbb{Z})^n$. For each ideal $i \in \{0, ..., s\}$, we define

$$n_i := |\{j \in \{1, \ldots, n\} \mid \langle \pi_j(\mathcal{C}) \rangle = \langle p^i \rangle\}|,$$

where for each $j \in \{1, \ldots, n\}$, π_j is the projection onto the j-th coordinate. We call $(n_0, \ldots, n_s)$ the support subtype of $\mathcal{C}$.

Lemma 18. Let $\mathcal{C}$ be an $\mathbb{Z}/p^s\mathbb{Z}$-submodule of $(\mathbb{Z}/p^s\mathbb{Z})^n$ of support subtype $(n_0, \ldots, n_s)$. Then

$$\overline{w}_L(\mathcal{C}) = \begin{cases} \dfrac{1}{4p^s}\left(p^{2s}(n-n_s) - \displaystyle\sum_{i=0}^{s-1} p^{2i} n_i\right) & \text{if } p \text{ is odd,} \\ 2^{s-2}(n-n_s) & \text{if } p = 2. \end{cases}$$

Theorem 19. Let $\mathcal{C}$ be a $\mathbb{Z}/p^s\mathbb{Z}$-submodule of $(\mathbb{Z}/p^s\mathbb{Z})^n$ and let $\mathcal{C}'$ be a non-trivial subcode of $\mathcal{C}$. Then

$$d_L(\mathcal{C}) \leq |\mathcal{C}'|(|\mathcal{C}'|-1)^{-1}\overline{w}_L(\mathcal{C}'). \tag{2}$$

Definition 20. Let p be a prime and let s be a positive integer. Define

$$A(p, s, i) := \begin{cases} \dfrac{p^{s-i}(p^i+1)}{4} & \text{if } p \text{ is odd,} \\ \dfrac{2^{s-2+i}}{2^i-1} & \text{if } p = 2. \end{cases}$$

With respect to this notation, the Chiang-Wolf bound of Proposition 13 is given by:

$$d_L(\mathcal{C}) \leq \lfloor A(p, s, s) \rfloor (n - k_1 + 1).$$

Theorem 21. Let $\mathcal{C}$ be a $\mathbb{Z}/p^s\mathbb{Z}$-submodule of $(\mathbb{Z}/p^s\mathbb{Z})^n$. Let $\ell \in \{1, \ldots, s\}$ such that there exists $y \in \mathcal{C}$ satisfying $w_H(y) = d_H(\langle y \rangle)$ and $y \in \langle p^{s-\ell} \rangle$. Then

$$d_L(\mathcal{C}) \leq A(p, s, \ell) d_H(\mathcal{C}).$$

For any linear code $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ there exist words in $\mathcal{C} \cap \langle p^{s-1} \rangle$ of Hamming weight $d_H(\mathcal{C})$, so certainly the hypothesis of Theorem 21 holds with $\ell = 1$.

Combining Proposition 7 and Theorem 19 we get the following bound.

Corollary 22. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a linear code of rank K. Then

$$d_L(\mathcal{C}) \leq \lfloor A(p, s, 1) \rfloor (n - K + 1). \tag{3}$$

Let $\ell \in \{1, \ldots, s\}$ such that there exists $y \in \mathcal{C}$ satisfying $w_H(y) = d_H(\langle y \rangle)$ and $y \in \langle p^{s-\ell} \rangle$. Then

$$d_L(\mathcal{C}) \leq \lfloor A(p, s, \ell) \rfloor (n - K + 1).$$

Corollary 23. Let $\mathcal{C}$ be a $\mathbb{Z}/p^s\mathbb{Z}$-submodule of $(\mathbb{Z}/p^s\mathbb{Z})^n$ of rank K. Then

$$\left\lfloor \frac{d_L(\mathcal{C}) - 1}{A(p, s, 1)} \right\rfloor \leq n - K.$$

Example 24. Let $\mathcal{C} = \langle (1, 2, 1, 3) \rangle \subset (\mathbb{Z}/5\mathbb{Z})^4$. Then $d_L(\mathcal{C}) = 6$. $\mathcal{C}$ meets the bound of Proposition 13 and Corollary 23.

Example 25. Let $\mathcal{C} = \langle (0, 1, 1), (2, 0, 0), (0, 0, 2) \rangle \subset (\mathbb{Z}/4\mathbb{Z})^3$. $\mathcal{C}$ attains the bound of Corollary 23 and does not attain the bound of Proposition 13.

4.1 Characterization

Lee-equidistant codes satisfy (1) with equality. The work of Wood (2001) and Dyshko (2019) yields the following.

Corollary 26. Let $\mathcal{C}$ be a constant weight submodule of $(\mathbb{Z}/p^s\mathbb{Z})^n$ of rank K with generator matrix G. Let U be the collection of orbits of $(\mathbb{Z}/p^s\mathbb{Z})^K$ under the action of $\{1, -1\}$. Then one of the following holds:

(1) $s = 1$ and a representative of each member of U appears as a column G with the same multiplicity,

(2) $p = 2$ and every member of $\mathbb{Z}/p^s\mathbb{Z}^K$ appears with the same multiplicity as a column of G,
(3) $K \leq 2$.

Setting $p = 2$ and $\ell = 1$ in Corollary 22 gives the same bound as $p = 2$ in Corollary 11, characterized in Theorem 17. We thus focus on the case p odd.

Proposition 27. Let p be an odd prime. Let $\mathcal{C} \subset (\mathbb{Z}/p^s\mathbb{Z})^n$ have rank K. If $\mathcal{C}$ meets the bound (3) of Corollary 22 then $n \leq p+1$ and either

$$K = n - p + 2 \leq 3 \text{ and } d_L(\mathcal{C}) = \frac{p^{s-1}(p^2-1)}{4}, \text{ or}$$

$$K = n + 1 - \frac{p-1}{2} \leq \frac{p+5}{2} \text{ and } d_L(\mathcal{C}) = \frac{p^{s-1}(p^2-1)}{8}.$$

Theorem 28. The density of optimal linear codes $\mathcal{C} \subset (\mathbb{Z}/p^s\mathbb{Z})^n$ of rank K for the bound of Corollary 22 with $\ell = 1$ is 0 as either $n \longrightarrow \infty$ or $p \longrightarrow \infty$.

5. CHARACTERIZING LEE-EQUIDISTANT CODES

Theorem 29. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a minimal-length linear Lee-equidistant code of rank $K = 1$ and minimum Lee distance w. Let i be the positive integer such that $k_i = 1$, then $\mathcal{C}$ has support subtype $(0, \ldots, 0, n_{i-1}, \ldots, n_{s-1}, 0)$ where, for $1 \leq j \leq s-1$,

$$w = \frac{p+1}{4} p^{s-1} n_{i-1} \text{ and } n_{i-1}(p-1) = p^{j-i+2} n_j.$$

As we have seen in Theorem 29 we have that $p^{s-i+1} \mid n_{i-1}$. Since the socle of $\mathcal{C}$ can be identified with a code over $\mathbb{F}_p$, from Corollary 26 we have that $n_{i-1} = \frac{p-1}{4}a$, for some $a \in \mathbb{N}$. With this we can exactly determine the support subtype of a smallest-length linear Lee-equidistant code.

Corollary 30. Let $\mathcal{C} \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a minimal-length linear Lee-equidistant code of rank $K = 1$ and minimum Lee distance w. Let i be the integer such that $k_i = 1$, then $\mathcal{C}$ has support subtype $(0, \ldots, 0, n_{i-1}, \ldots, n_{s-1}, 0)$ where, for all $j \in \{i, \ldots, s-1\}$,

$$w = p^{2s-i}\frac{p^2-1}{8}, n_{i-1} = p^{s-i+1}\frac{p-1}{2}, n_j = p^{s-j-1}\frac{(p-1)^2}{2}.$$

Corollary 31. Let $\mathcal{C} = \langle g_1, g_2 \rangle \subseteq (\mathbb{Z}/p^s\mathbb{Z})^n$ be a minimal-length linear Lee-equidistant code of rank $K = 2$ and minimum Lee distance w, with $g_1 \in \langle p^{i-1} \rangle$ and $g_2 \in \langle p^{s-1} \rangle$. Then $\mathcal{C}$ has support subtype of the form $(0, \ldots, 0, n_{i-1}, \ldots, n_{s-1}, 0)$ and $\langle g_1 \rangle, \langle g_2 \rangle$ have respective support subtypes

$$(0, \ldots, 0, n_{i-1}, \ldots, n_{s-2}, n_{s-1}^{(1)}, n_s^{(1)}),$$
$$(0, \ldots, 0, n_{s-1}^{(2)}, n_s^{(2)}),$$

where,

$$w = p^{2s-i}\frac{p^2-1}{8}, n_{i-1} = p^{s-i+1}\frac{p-1}{2}, n_s^{(1)} = \frac{p-1}{2},$$
$$n_s^{(2)} = p^{s-i}\frac{p-1}{2}, n_\ell^{(1)} = p^{s-\ell-1}\frac{(p-1)^2}{2},$$

for all $\ell \in \{i, \ldots, s-1\}$.

For each $\ell \in \{i, \ldots, s-1\}$, let U_ℓ be the set of all elements in $\langle p^\ell \rangle \setminus \langle p^{\ell+1} \rangle$ up to ± 1. We denote by u_ℓ the length $(p-1)\alpha_\ell$ tuple consisting of $p-1$ repetitions of each element in U_ℓ. We denote by $\tilde{u}_{i-1}$ the tuple of length $p\alpha_{i-1}$ consisting of p repetitions of all elements in U_{i-1}. Let x be the length $\frac{p-1}{2}$ tuple of all elements in U_{s-1}, let $y = (x, -x)$, and let $z = (0, y)$. Let a be $p^{s-i}\frac{p-1}{2}$ copies of z, b be $\left(p^{s-i}-1\right)\frac{p-1}{2}$ copies of y, and let c be $\frac{p-1}{2}$ copies of x.

Theorem 32. The matrix

$$G = \begin{pmatrix} \tilde{u}_{i-1} & u_i & \cdots & u_{s-1} & \mathbf{0} \\ a & & b & & c \end{pmatrix}$$

generates a Lee-equidistant code over $\mathbb{Z}/p^s\mathbb{Z}$ with $k_i = 1, k_s = 1$.

Example 33. Let us consider the case $\mathbb{Z}/27\mathbb{Z}$ and $k_1 = 1, k_3 = 1$. Let

$$\tilde{u}_0 = (1,1,1,2,2,2,4,4,4,\ldots,13,13,13),$$

i.e., the tuple consisting of 3 repetitions of all elements in

$$\{1,2,4,5,7,8,10,11,13\},$$

let $u_1 = (3,3,6,6,12,12)$ and $u_2 = (9,9)$. Let a be the tuple consisting of 9 repetitions of $(0,9,18)$ and b be the tuple consisting of 4 repetitions of $(9,18)$. Then the matrix

$$G = \begin{pmatrix} \tilde{u}_0 & u_1 & u_2 & 0 \\ a & b & 9 & \end{pmatrix}$$

generates a Lee-equidistant code.

REFERENCES

Alderson, T.L. and Huntemann, S. (2013). On maximum Lee distance codes. *Journal of Discrete Mathematics*, 2013.

Chiang, J.C.Y. and Wolf, J.K. (1971). On channels and codes for the Lee metric. *Information and Control*, 19(2), 159–173.

Dougherty, S.T. (2017). *Algebraic coding theory over finite commutative rings*. Springer.

Dougherty, S.T. and Shiromoto, K. (2000). MDR codes over $\mathbb{Z}_k$. *IEEE Transactions on Information Theory*, 46(1), 265–269.

Dyshko, S. (2019). The extension theorem for Lee and Euclidean weight codes over integer residue rings. *Designs, Codes and Cryptography*, 87(6), 1253–1269.

Honold, T. and Landjev, I. (2000). Linear codes over finite chain rings. *The Electronic Journal of Combinatorics*, 7, R11–R11.

Shiromoto, K. (2000). Singleton bounds for codes over finite rings. *Journal of Algebraic Combinatorics*, 12(1), 95–99.

Wood, J.A. (2001). The structure of linear codes of constant weight. *Trans. American Math. Soc.*, 354(3), 1007–1026.

Wyner, A.D. and Graham, R.L. (1968). An upper bound on minimum distance for a k-ary code. *Inf. Control.*, 13(1), 46–52.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Combinatorial tools for the study of flag codes ⋆

C. Alonso-González * **M.A. Navarro-Pérez** **

* *Department of Mathematics, University of Alicante, Alicante, Spain (e-mail: clementa.alonso@ua.es).*
** *Department of Mathematics, University of Alicante, Alicante, Spain (e-mail: miguelangel.np@ua.es)*

Abstract: In network coding, a *flag code* is a collection of *flags*, that is, sequences of nested subspaces of $\mathbb{F}_q^n$, being $\mathbb{F}_q$ the finite field with q elements. If the sequence of subspace dimensions is $(1, 2, \ldots, n-1)$, we speak about full flag codes. The familiy of flag codes was first introduced in Liebhold et al. (2018). In this work we present some combinatorial tools coming from the classical theory of partitions that can be naturally associated with full flag codes in order to extract relevant information about them. In particular, we state a combinatorial characterization of those full flag codes that attain the maximum possible distance.

Keywords: Flag codes, flag codistance, Ferrers diagrams, integer partitions.

1. INTRODUCTION

In Ahlswede et al. (2000), random network coding was introduced as the most efficient way to send data across a non-coherent network with multiple sources and sinks. However, it is very susceptible to error propagation and erasures. To solve this problem, in Koetter et al. (2008) the authors propose just considering subspaces of $\mathbb{F}_q^n$ as the codewords of *subspace codes*. Since this seminal paper, much research has been made to construct large subspace codes as well as to determine their properties. In case all the subspaces in the code have the same dimension, we have *constant dimension codes*. To have an overview of the most important works in this subject, consult Trautmann et al. (2018) and references therein.

In Liebhold et al. (2018) the authors developed techniques for the use in network coding of constant type *flags*, that is, sequences of nested subspaces of prescribed dimensions. In this context, collections of flags are denominated *flag codes* and they generalize constant dimension codes. The recent works Alonso-González et al. (2020); Alonso-González et al. (2021a); Kurz (2021) are dedicated to this topic.

A way to naturally associate constant dimension codes with a flag code is by considering all the subspaces of a given dimension of all the flags in a flag code. In this way we obtain a *projected code*. In the study of flag codes, one of the principal proplems is the determination of the relationship between the parameters of a flag code and the ones of its projected codes. In Alonso-González et al. (2021b); Alonso-González et al. (2020); Alonso-González et al. (2021a); Navarro-Pérez et al. (2021) this question has been undertaken for the family of flag codes attaining the maximum distance (*optimum distance flag codes*) whereas in Alonso-González and Navarro-Pérez (2020), the authors define *consistent* flag codes as precisely those whose projected codes completely determine their parameters. In this work we contribute to advance in this question from an innovative combinatorial perspective.

One of the main difficulties when investigating the parameters of a flag code relies on the definition of the distance between flags provided that it is obtained as the sum of their subspace distances. This causes that we can attain a flag distance value by many different combinations. To capture such a variability, in Alonso-González et al. (2021), the authors introduce the notion of *distance vector* (associated to a given distance value). Here, we draw distance vectors in the *distance support* to obtain the so-called *distance paths*. This geometrical representation allows us to introduce the *codistance* of the flag code (the complement of the distance) and hence, naturally associate to a flag code different combinatorial objects coming from the classical theory of partitions that result very convenient for our purposes. The results presented in this work summarize the ones appearing in Alonso-González and Navarro-Pérez (2021).

2. SOME PRELIMINAIRES

2.1 Partitions and Ferrers diagrams

Let us first fix some notation on integer partitions and their representation by Ferrers diagrams. Given a positive integer s, a *partition* of s is a sequence of non-increasing positive integers $\lambda = (\lambda_1, \ldots, \lambda_m)$ such that $\lambda_1 + \cdots + \lambda_m = s$. Each value λ_i is called a *part* of λ and we say that m is the *length* of λ.

Ferrers diagrams allow us to give geometrical representations of partitions and to extract relevant properties about them in some cases.

Given a partition $\lambda = (\lambda_1, \ldots, \lambda_m)$, its *associated Ferrers diagram* $\mathfrak{F}_\lambda$ is constructed by stacking right-justified m

⋆ The authors receive financial support from Ministerio de Ciencia e Innovación (PID2019-108668GB-I00).

rows of dots, where the number of dots in the i-th row is λ_i. The dot at the top right position is called the *corner* of the Ferrers diagram.

2.2 Flags and flag distance

Throughout the paper q will denote a fixed prime power and k, n two integers with $1 \leq k < n$. Consider $\mathbb{F}_q$ the finite field with q elements and denote by $\mathcal{G}_q(k,n)$ the *Grassmannian*, that is, the set of k-dimensional subspaces of $\mathbb{F}_q^n$. This set can be equipped with the so-called *injection distance*: given two subspaces $\mathcal{U}, \mathcal{V} \in \mathcal{G}_q(k,n)$, it is defined as

$$d_I(\mathcal{U},\mathcal{V}) = k - \dim(\mathcal{U} \cap \mathcal{V}). \tag{1}$$

Using this distance, we can define a *constant dimension code* $\mathcal{C}$ of length n and dimension k as a nonempty subset of $\mathcal{G}_q(k,n)$. The *minimum distance* of $\mathcal{C}$ is defined as

$$d_I(\mathcal{C}) = \min\{d_I(\mathcal{U},\mathcal{V}) \mid \mathcal{U},\mathcal{V} \in \mathcal{C},\ \mathcal{U} \neq \mathcal{V}\}$$

whenever $|\mathcal{C}| \geq 2$. In case $|\mathcal{C}| = 1$, we put $d_I(\mathcal{C}) = 0$.

See Trautmann et al. (2018) and the references therein for more information on this class of codes.

The concept of constant dimension code can be extended when considering flags of constant type on $\mathbb{F}_q^n$, that is, sequences of nested subspaces of $\mathbb{F}_q^n$ where the list of corresponding dimensions is fixed. The use of flags in network coding as a generalization of constant dimension codes was first proposed in Liebhold et al. (2018). Let us recall some basic background on flag codes.

A *flag* $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_r)$ on $\mathbb{F}_q^n$ is a sequence of nested $\mathbb{F}_q$-vector subspaces $\{0\} \subsetneq \mathcal{F}_1 \subsetneq \cdots \subsetneq \mathcal{F}_r \subsetneq \mathbb{F}_q^n$. The vector $(\dim(\mathcal{F}_1), \ldots, \dim(\mathcal{F}_r))$ is called the *type* of $\mathcal{F}$ and $\mathcal{F}_i$ is the *i-th subspace* of $\mathcal{F}$. In particular, if the type vector is $(1, 2, \ldots, n-1)$, we say that $\mathcal{F}$ is a *full flag*.

The set of all the flags on $\mathbb{F}_q^n$ of a fixed type vector $(t_1, \ldots, t_r)$ is said to be the *flag variety* $\mathcal{F}_q((t_1, \ldots, t_r), n)$ and, for every $i = 1, \ldots, r$, we define the *i-projection* as the map $p_i : \mathcal{F}_q((t_1, \ldots, t_r), n) \to \mathcal{G}_q(t_i, n)$, given by $p_i((\mathcal{F}_1, \ldots, \mathcal{F}_r)) = \mathcal{F}_i$.

The flag variety $\mathcal{F}_q((t_1, \ldots, t_r), n)$ is a metric space: given two flags $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_r)$ and $\mathcal{F}' = (\mathcal{F}'_1, \ldots, \mathcal{F}'_r)$ in $\mathcal{F}_q((t_1, \ldots, t_r), n)$, the *(injection) flag distance* between them is the value

$$d_f(\mathcal{F},\mathcal{F}') = \sum_{i=1}^{r} d_I(\mathcal{F}_i, \mathcal{F}'_i). \tag{2}$$

A *flag code* of type $(t_1, \ldots, t_r)$ on $\mathbb{F}_q^n$ is a nonempty subset $\mathcal{C} \subseteq \mathcal{F}_q((t_1, \ldots, t_r), n)$. Its *minimum distance* is given by

$$d_f(\mathcal{C}) = \min\{d_f(\mathcal{F},\mathcal{F}') \mid \mathcal{F},\mathcal{F}' \in \mathcal{C}, \mathcal{F} \neq \mathcal{F}'\}.$$

when $|\mathcal{C}| \geq 2$. If $|\mathcal{C}| = 1$, we put $d_f(\mathcal{C}) = 0$. The *i-projected code* of $\mathcal{C}$ is the set

$$\mathcal{C}_i = \{p_i(\mathcal{F}) \mid \mathcal{F} \in \mathcal{C}\} \subseteq \mathcal{G}_q(t_i, n).$$

Remark 1. Note that the i-projected code $\mathcal{C}_i$ of $\mathcal{C}$ is a constant dimension code in the Grassmannian $\mathcal{G}_q(t_i, n)$ closely related to the flag code $\mathcal{C}$. Moreover, the cardinality of $|\mathcal{C}_i|$ always satisfies $|\mathcal{C}_i| \leq |\mathcal{C}|$, whereas there is not a clear relationship concerning the distance. In fact, we can have $d_f(\mathcal{C}) > d_I(\mathcal{C}_i)$, $d_f(\mathcal{C}) = d_I(\mathcal{C}_i)$ or even $d_f(\mathcal{C}) < d_I(\mathcal{C}_i)$. So, the problem of obtaining the parameters of a flag code from the ones of its projected codes and conversely is a central one in the study of flag codes.

3. COMBINATORIAL TOOLS

From now on we work just with full flags. For each dimension $0 \leq i \leq n$, we define *the distance support* $\mathrm{S}(i,n)$ of $\mathbb{Z}^2$ by

$$\mathrm{S}(i,n) = \{i\} \times \{0, 1, \ldots, \min\{i, (n-i)\}\}. \tag{3}$$

We extend it to the full flag variety as follows: the *distance support* of the full flag variety on $\mathbb{F}_q^n$ is the set

$$\mathrm{S}(n) = \bigcup_{i=0}^{n} \mathrm{S}(i,n) \subset \mathbb{Z}^2. \tag{4}$$

3.1 Distance paths

If we consider two full flags $\mathcal{F}, \mathcal{F}'$ on $\mathbb{F}_q^n$, their flag distance can be geometrically represented by means of a collection of $n+1$ points in the distance support $\mathrm{S}(n)$, each one of them in a different column $\mathrm{S}(i,n)$, then we can define their *distance path* $\Gamma(\mathcal{F},\mathcal{F}')$ as the directed polygonal path whose vertices are the points $(i, d_I(\mathcal{F}_i, \mathcal{F}'_i))$ for every $0 \leq i \leq n$.

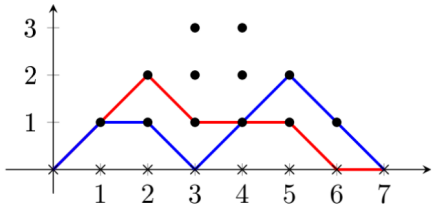


Fig. 1. Two distance paths in S(7).

Similarly, the set of distance paths of a given flag code $\mathcal{C}$ is

$$\Gamma(\mathcal{C}) = \{\Gamma(\mathcal{F},\mathcal{F}') \mid \mathcal{F},\mathcal{F}' \in \mathcal{C},\ \mathcal{F} \neq \mathcal{F}'\}.$$

3.2 Flag codistance

The simple idea of drawing the distance between flags by means of a distance path in a suitable distance support allows us to pay attention to the complementary parameter to the flag distance. If D^n is the maximum possible distance between flags in the full flag variety, given a flag distance value d, i.e., an integer such that $0 \leq d \leq D^n$, we define its *(injection flag) codistance* as the value $\bar{d} = D^n - d$. Similarly, given a full flag code $\mathcal{C}$ on $\mathbb{F}_q^n$, we define its associated *codistance* as the value $\bar{d}_f(\mathcal{C}) = D^n - d_f(\mathcal{C})$.

3.3 Ferrers diagram frame

Now, fixed a positive integer n, we consider the enriched distance support $\hat{\mathrm{S}}(n)$ by adding auxiliary red points and making a rotation around the point $(n, 0)$ as below.

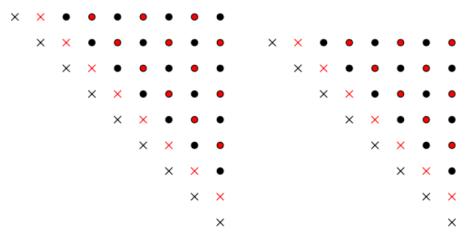


Fig. 2. Rotated enriched supports $\hat{\mathrm{S}}(8)$ and $\hat{\mathrm{S}}(7)$.

If we eliminate the crossed dots, we obtain a Ferrers diagram frame $\mathrm{FF}(n)$. Here we can stablish a one-to-one correspondence between the set of distance paths (associated to a given value of the flag distance) and certain partitions that correspond to the set of circle black points cointained in a suitable *Ferrer subdiagram* of $\mathrm{FF}(n)$.

3.4 Ferrers subdiagrams and embedded partitions

Every Ferrers diagram contained in $\mathrm{FF}(n)$ is said to be a *Ferrers subdiagram*. We also say that the partition $\lambda = (\lambda_1, \ldots, \lambda_m)$ of the integer $\sum_{i=1}^{m} \lambda_i$ is an *embedded partition* on $\mathrm{FF}(n)$ if

(1) $1 \leq m \leq n-1$ and,
(2) for every $1 \leq i \leq m$, it holds $\lambda_i \leq n - i$.

We denote by $\mathfrak{F}_\lambda$ the Ferrers subdiagram associated to an embedded partition λ and also consider the *empty Ferrers subdiagram* $\mathfrak{F}_0$, associated to the *null embedded partition* $\lambda = (0)$.

Two Ferrers subdiagrams of $\mathrm{FF}(n)$ are said to be *distance-equivalent* if they have the same *underlying black diagram*, that is, they contain the same set of black dots. Analogously, two embedded partitions λ and λ' are said to be *distance-equivalent* if their associated Ferrers subdiagrams $\mathfrak{F}_\lambda$ and $\mathfrak{F}_{\lambda'}$ are.

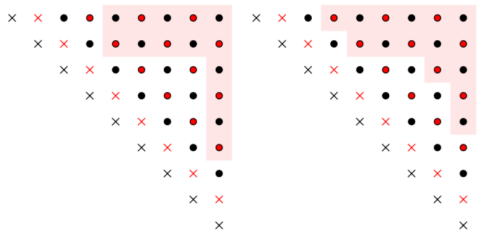


Fig. 3. Distance-equivalent Ferrers subdiagrams in $\mathrm{FF}(8)$.

We can show that a distance path with distance d, determines a set of Ferrers subdiagrams, all of them being distance-equivalent. In fact, each of those Ferrers diagrams is associated to an underlying partition to which we can associate a distinguished value: given $\lambda = (\lambda_1, \ldots, \lambda_m)$ an embedded partition in $\mathrm{FF}(n)$, we define its *underlying distribution* as the vector

$$U_\lambda = \begin{cases} \left(\left\lceil \frac{\lambda_1}{2} \right\rceil, \left\lfloor \frac{\lambda_2}{2} \right\rfloor, \left\lceil \frac{\lambda_3}{2} \right\rceil, \ldots \right) & \text{if } n \text{ is even,} \\ \left(\left\lfloor \frac{\lambda_1}{2} \right\rfloor, \left\lceil \frac{\lambda_2}{2} \right\rceil, \left\lfloor \frac{\lambda_3}{2} \right\rfloor, \ldots \right) & \text{if } n \text{ is odd.} \end{cases}$$

We denote by u_λ the sum of the components of U_λ. With this notation, the next result holds.

Theorem 2. Let Γ_d be a distance path associated to the distance value d. If $\mathfrak{F}_\lambda$ is a Ferrers subdiagram determined by Γ_d, then it holds

$$\bar{d} = D^n - d = u_\lambda.$$

3.5 Codistance splittings

We introduce a new concept that relates embedded partitions and codistance. Given $\lambda = (\lambda_1, \ldots, \lambda_m)$ an embedded partition in $\mathrm{FF}(n)$, we say that its underlying distribution U_λ *splits* the value u_λ defined above, or that it is an *splitting* of u_λ. This value u_λ is common for $\mathfrak{F}_\lambda$ and all its distance-equivalent Ferrers subdiagrams.

Finally, the next result provides the bridge to translate the information given by distance paths to the embedded partitions level and conversely.

Theorem 3. Let $n \geq 2$ be an integer and $0 \leq d \leq D^n$ a flag distance value. Then there is a bijection between the set of distance paths of distance d in $\mathrm{S}(n)$ and the set of splittings of the codistance $\bar{d} = D^n - d$.

4. DERIVED RESULTS FOR FLAG CODES

The previously introduced combinatorial tools can be applied to establish connections between the parameters of a given full flag code and the ones of its projected codes. We summarize here some of the most important consequences. See Alonso-González and Navarro-Pérez (2021) for more details.

Theorem 4. Consider a full flag code $\mathcal{C}$ on $\mathbb{F}_q^n$ with codistance $\bar{d}_f(\mathcal{C})$ and take a dimension $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$. If the codistance satisfies

$$\bar{d}_f(\mathcal{C}) < \left\lceil \frac{i(n-i)}{2} \right\rceil, \tag{5}$$

then $|\mathcal{C}| = |\mathcal{C}_i| = \cdots = |\mathcal{C}_{n-i}|$.

Theorem 5. Let $\mathcal{C}$ be a full flag code on $\mathbb{F}_q^n$ with associated codistance $\bar{d}_f(\mathcal{C})$. Take a dimension $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$ and consider an integer $0 \leq r \leq i$. Hence, whenever

$$\bar{d}_f(\mathcal{C}) < \left\lceil \frac{r(r+n-2i)}{2} \right\rceil, \tag{6}$$

then $|\mathcal{C}_i| = |\mathcal{C}|$ and $d_I(\mathcal{C}_i) > i - r$.

Theorem 6. Let $\mathcal{C}$ be a full flag code on $\mathbb{F}_q^n$ and take $1 \leq i \leq \lfloor \frac{n}{2} \rfloor$.

(1) If $|\mathcal{C}_i| = |\mathcal{C}|$, then

$$d_I(\mathcal{C}_i)^2 \leq d_f(\mathcal{C}) \leq D^n - \left\lceil \frac{(i - d_I(\mathcal{C}_i))(n - i - d_I(\mathcal{C}_i))}{2} \right\rceil$$

(2) If $|\mathcal{C}_i| < |\mathcal{C}|$, then

$$0 \leq d_f(\mathcal{C}) \leq D^n - \left\lceil \frac{i(n-i)}{2} \right\rceil.$$

To finish, we state a combinatorial characterization of full flag codes attaining the maximum possible distance:

Theorem 7. Let $\mathcal{C}$ be a full flag code on $\mathbb{F}_q^n$. They are equivalent:

(1) $d_f(\mathcal{C}) = D^n$ (or $\bar{d}_f(\mathcal{C}) = 0$).

(2) The set $\Gamma(\mathcal{C})$ consists of the only distance path passing either through the point $(\frac{n}{2}, \frac{n}{2})$, if n is even, or through the points $(\lfloor \frac{n}{2} \rfloor, \lfloor \frac{n}{2} \rfloor)$ and $(\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor)$, if n is odd.

(3) The set of Ferrers subdiagrams associated to $\mathcal{C}$ is

$$\mathfrak{F}(\mathcal{C}) = \begin{cases} \{\mathfrak{F}_0\} & \text{if } n \text{ is even or} \\ \{\mathfrak{F}_0, \mathfrak{F}_{(1)}\} & \text{if } n \text{ is odd.} \end{cases}$$

REFERENCES

R. Ahlswede, N. Cai, R. Li and R. W. Yeung, Network Informatin Flow. *IEEE Transactions on Information Theory*, Vol. 46 (2000), 1204-1216.

C. Alonso-González and M. A. Navarro-Pérez, A combinatorial approach to flag codes, *https://arxiv.org/abs/2111.15388* (preprint).

C. Alonso-González and M. A. Navarro-Pérez, Consistent Flag Codes, *Mathematics*, Vol. 8(12) (2020), 2243.

C. Alonso-González, M. A. Navarro-Pérez and X. Soler-Escrivà, An orbital construction of Optimum Distance Flag Codes. *Finite Fields and Their Applications*, Vol. 73 (2021), 101861.

C. Alonso-González, M. A. Navarro-Pérez and X. Soler-Escrivà, Optimum Distance Flag Codes from Spreads via Perfect Matchings in Graphs, *Journal of Algebraic Combinatorics*, Vol. 54 (2021), 1279–1297.

C. Alonso-González, M. A. Navarro-Pérez and X. Soler-Escrivà, Flag Codes from Planar Spreads in Network Coding, *Finite Fields and Their Applications*, Vol. 68 (2020), 101745.

C. Alonso-González, M. A. Navarro-Pérez and X. Soler-Escrivà, Flag Codes: Distance Vectors and Cardinality Bounds, *https://arxiv.org/abs/2111.00910* (preprint).

G. E. Andrews, Generalizations of the Durfee square, *Journal of the London Mathematical Society*, Vol. 3 (2) (1971), 563-570.

G. E. Andrews, The Theory of Partitions, *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, (1984).

R. Koetter and F. Kschischang, Coding for Errors and Erasures in Random Network Coding, *IEEE Transactions on Information Theory*, Vol. 54 (2008), 3579-3591.

S. Kurz, Bounds for Flag Codes, *Designs, Codes and Cryptography*, Vol. 89 (2021), 2759–2785.

D. Liebhold, G. Nebe and A. Vazquez-Castro, Network Coding with Flags, *Designs, Codes and Cryptography*, Vol. 86 (2) (2018), 269-284.

M. A. Navarro-Pérez and X. Soler-Escrivà, Flag Codes of Maximum Distance and Constructions using Singer Groups, *https://arxiv.org/abs/2109.00270* (preprint).

A.-L. Trautmann and J. Rosenthal, Constructions of Constant Dimension Codes, in: M. Greferath *et al.* (Eds.), *Network Coding and Subspace Designs*, E-Springer International Publishing AG, 2018, pp. 25-42.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Average Turnpike Property

Martín Hernández * **Rodrigo Lecaros** **
Sebastián Zamorano ***

* *Friedrich-Alexander-Universität Erlangen-Nürnberg. Erlangen, Germany. (e-mail: martin.hernandez@fau.de).*
** *Universidad Técnica Federico Santa María. Santiago, Chile. (e-mail:rodrigo.lecaros@usm.cl)*
*** *University of Santiago of Chile. Santiago, Chile. (e-mail: sebastian.zamorano@usach.cl)*

Abstract: This paper studies the integral turnpike and turnpike in average for a class of random ordinary differential equations. We prove that, under suitable assumptions on the matrices that define the system, the optimal solutions for an optimal distributed control tracking problem remain, in an average sense, sufficiently close to the associated random stationary optimal solution for most of the time horizon.

Keywords: Turnpike property, exponential turnpike property, random ODE, long time behavior, model predictive control.

1. INTRODUCTION

We consider an optimal control distributed tracking-type problem of linear ordinary differential equations with random coefficients. This kind of differential equation is the stochastic counterpart of deterministic differential equations. The term random differential equations, in general, refers to differential equations with random coefficients, having either deterministic or random inhomogeneous parts and initial conditions.

Differential equations with random coefficients have been studied and used for various engineering and science problems. The latter is because the solution of a dynamic system is a function that depends on the parameters which constitute the system. These parameters are experimentally determined and are usually the mean value of a set of experimental observations. However, the observations might be measured with errors due to the conditions' variability, uncertainties, or lack of knowledge. Therefore, a suitable approach to analysis would be to consider systems with random variables as coefficients. We can mention the earlier work on this area Bergmann (1946) where the author studied the propagation of high-frequency sound waves in the atmosphere of randomly varying refraction index. We refer to the books Soong (1973, 1981) for a complete study of this kind of equations and interesting applications in science, engineering, physics, and biomedical systems, among others.

On the other hand, in the context of optimal control, in Porretta and Zuazua (2013) the authors studied the concept of the turnpike for the solutions of an optimal control problem subject to ordinary differential equations without randomness. The turnpike property, roughly speaking, describes that the optimal evolutionary solution is made of three arcs. The first and the last arcs are transient short-time, and the middle piece is a long-time arc remaining exponentially close to the optimal steady-state of the corresponding stationary optimal control problem. This concept was formulated in the earlier work Dorfman et al. (1987), in the context of the econometric field.

Motivated by the previous considerations, we will investigate if any connection exists between the average, with respect to the random variable, of an optimal solution of a certain optimal control problem for an ordinary differential equation with random coefficients, with the corresponding stationary random problem. Specifically, we will analyze the *turnpike phenomenon* for a class of random differential equations, which is important to understand the behavior of solutions to optimal control problems on large time horizons. In the context of stochastic differential equations, this property has been the focus of recent interest Sun et al. (2022). However, to the best of our knowledge, it is the first time that the turnpike property has been studied for ordinary differential equations with random coefficients.

Stating things more mathematically, in this paper we consider a probability space $(\Omega, \mathcal{F}, \mu)$ and three random matrices $A, C \in C^0(\Omega, \mathcal{L}(\mathbb{R}^n))$ and $B \in C^0(\Omega, \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n))$, constant in time, which will represent the random coefficients of the equation, the random observation and the random control, respectively. We assume that the joint probability distribution of matrices A, B and C is specified. We consider the following optimal control problem

$$\min_{u\in L^2(0,T;\mathbb{R}^m)} \left\{ J^T(u) = \frac{1}{2}\left(\int_0^T \|u(t)\|^2_{\mathbb{R}^m} dt + \int_0^T \|C(\cdot)x(t,\cdot) - z\|^2_{L^2(\Omega,\mathbb{R}^n)} dt \right) \right\}, \tag{1}$$

⋆ The work of R. Lecaros was partially supported by FONDECYT Grant 11180874. S. Zamorano was partially supported by the ANID-PAI Convocatoria Nacional Subvención a la Instalación en la Academia Convocatoria 2019 PAI77190106.
⋆⋆2020 Mathematics Subject Classification. Primary: 49K15, 49K40; Secondary: 49K45, 93D20.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

subject to x solving the following evolutionary problem

$$\begin{cases} x_t + A(\omega)x = B(\omega)u & t \in (0,T), \\ x(0,\omega) = x_0(\omega), \end{cases} \tag{2}$$

where $z \in \mathbb{R}^n$ is a fixed target. Here $x = x(t,\omega) \in \mathbb{R}^n$ represents the state and $u(t) \in \mathbb{R}^m$ the control of the system, respectively.

The first aim of this paper it is prove that when the time-horizon goes to infinity, the optimal pair (u^T, x^T) of (1)-(2) converges in an averaged sense to $(\overline{u}, \overline{x})$ in $\mathbb{R}^m \times L^2(\Omega, \mathbb{R}^n)$, where $(\overline{u}, \overline{x})$ solves the associated stationary random optimal control problem

$$\min_{u \in \mathbb{R}^m} \left\{ J^s(u) = \frac{1}{2}\Big(\|u\|^2_{\mathbb{R}^m} + \|C(\cdot)x(\cdot) - z\|^2_{L^2(\Omega,\mathbb{R}^n)} \Big) \right\}, \tag{3}$$

subject to x solving the problem

$$A(\omega)x(\omega) = B(\omega)u, \tag{4}$$

where $z \in \mathbb{R}^n$ is the same target of the problem (1). That is, we will analyze the following limits, which are usually called integral turnpike property

$$\begin{aligned} &\frac{1}{T}\int_0^T x^T(t,\cdot)dt \to \overline{x}(\cdot) \quad \text{in } L^2(\Omega,\mathbb{R}^n), \\ &\frac{1}{T}\int_0^T u^T(t)dt \to \overline{u} \quad \text{in } \mathbb{R}^m. \end{aligned} \tag{5}$$

The second main result is to show an exponential turnpike property. Namely, we will prove the existence of two positive constants K and δ, independent on the time-horizon T, such that the solutions of the extremal equations (u^T, x^T, φ^T) remains exponentially close to the steady state solution, the so-called turnpike, for the majority of the time. That is,

$$\|x^T(t,\cdot) - \overline{x}(\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n)} + \|\varphi^T(t,\cdot) - \overline{\varphi}(\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n)} \leq K(e^{-\delta(T-t)} + e^{-\delta t}),$$

for every $t \in (0,T)$. Here, $(\varphi^T, \overline{\varphi})$ represents the characterization of minimizers via the first order necessary optimality conditions (the dual variables). In addition, as a consequence of the previous estimate, we prove an average exponential turnpike, that is

$$\|\mathbb{E}(x^T - \overline{x})\|_{\mathbb{R}^n} + \|\mathbb{E}(\varphi^T - \overline{\varphi})\|_{\mathbb{R}^n} + \|u^T - \overline{u}\|_{\mathbb{R}^m} \leq K(e^{-\delta(T-t)} + e^{-\delta t}),$$

for every $t \in (0,T)$ and where $\mathbb{E}(x^T)$ denotes the expected value of x^T and is given by

$$\mathbb{E}(x^T) = \int_\Omega x^T d\mu. \tag{6}$$

Let us mention that both results are based on stability assumptions for A, B, and C, which are the matrices that define the system. These assumptions are related to the existence of feedback operators in such a way that we can ensure an ellipticity-type condition. Besides, these hypotheses allow us to establish relevant observability inequalities, which play an essential role in the proof of our main results. We refer to Section 2 for a complete discussion on the subject. As a final remark, the C matrix must not be square. However, we will continue to assume $C \in C^0(\Omega, \mathcal{L}(\mathbb{R}^n))$ just for simplicity.

2. OPTIMAL CONTROL PROBLEMS

We consider the following random ODE

$$\begin{cases} x_t + A(\omega)x = B(\omega)u & t \in (0,T), \\ x(0,\omega) = x_0(\omega), \end{cases}$$

where $\omega \in \Omega$ corresponds the random parameter, $x = x(t,\omega) \in \mathbb{R}^n$ is the state of the system, the $n \times n$ matrix $A \in C^0(\Omega, \mathcal{L}(\mathbb{R}^n))$ governs its free dynamics, $u(t) \in \mathbb{R}^m$ is the control function which is assumed to be independent of the randomness and acts on the system through the control matrix $B \in C^0(\Omega, \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n))$ which is a $m \times n$ parameter dependent matrix. The initial datum $x_0(\omega)$ belongs to the space $L^2(\Omega, \mathbb{R}^n; \mu)$, which is defined below.

Let us define the space

$$L^2(\Omega,\mathbb{R}^n;\mu) := \Big\{ \omega \in \Omega \mapsto y(\omega) \in \mathbb{R}^n \text{ measurable } : \|y(\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n;\mu)} = \int_\Omega \|y(\omega)\|^2_{\mathbb{R}^n} d\mu(\omega) < \infty \Big\},$$

which is a Hilbert space endowed with the inner product

$$\langle x, y \rangle_{L^2(\Omega,\mathbb{R}^n;\mu)} = \int_\Omega \langle x(\omega), y(\omega) \rangle_{\mathbb{R}^n} d\mu(\omega), \quad \forall x, y \in L^2(\Omega,\mathbb{R}^n;\mu).$$

In what follows, we denote by $L^2(\Omega,\mathbb{R}^n) := L^2(\Omega,\mathbb{R}^n;\mu)$.

Additionally, we also assume that the matrices A and B are uniformly bounded with respect to ω.

Concerning the integrability of the solutions for (2), we have the next result, which can be consulted in Lohéac and Zuazua (2016).

Theorem 1. ((Lohéac and Zuazua, 2016, Corollary 2.2)).

Assume the map $\omega \mapsto (A(\omega), B(\omega))$ is continuous on Ω. Then, for every $x_0 \in L^2(\Omega,\mathbb{R}^n)$, every $u \in L^2_{loc}(\mathbb{R}^+,\mathbb{R}^m)$, and every $t \geq 0$, the solution x of (2) satisfies $x(t,\cdot) \in L^2(\Omega,\mathbb{R}^n)$. In addition, the solution x can be represented by

$$x(t,\omega) = e^{tA(\omega)}x_0(\omega) + \int_0^T e^{(t-s)A(\omega)}B(\omega)u(s)ds, \quad \forall \omega \in \Omega,\ t \in [0,T].$$

2.1 Evolutionary problem

Let us consider first the optimal control for the evolutionary problem (2) with initial datum independent of ω. That is,

$$\min_{u \in L^2(0,T;\mathbb{R}^m)} J^T(u), \tag{7}$$

where

$$J^T(u) = \frac{1}{2}\Big(\int_0^T \|u(t)\|^2_{\mathbb{R}^m} dt + \int_0^T \|C(\cdot)x(t,\cdot) - z\|^2_{L^2(\Omega,\mathbb{R}^n)} dt \Big),$$

subject to x solving the following evolutionary problem

$$\begin{cases} x_t + A(\omega)x = B(\omega)u & t \in (0,T), \\ x(0,\omega) = x_0, \end{cases} \tag{8}$$

where $z \in \mathbb{R}^n$ is a fixed target. The case of initial condition $x_0 \in L^2(\Omega,\mathbb{R}^n)$ does not lead to any essential new difficulty throughout the following. Thus, for sake of simplicity of the presentation, we only deal with the case where x_0 is independent of ω.

By using the direct method of the calculus of variations and noting that the solution x of (8) depends linearly and continuously on u, we obtain the existence and uniqueness result of the optimal control. Besides, the characterization of the control can be done using the Gateaux derivative of J^T. These results are included in the following theorem.

Theorem 2. There exists a unique solution $(u^T, x^T) \in L^2(0,T;\mathbb{R}^m) \times C^0([0,T];L^2(\Omega,\mathbb{R}^n))$ to the minimization problem (7)-(8), where x^T is the optimal state associated to the control u^T. In addition,

$$u^T(t) = -\int_\Omega B^*(\omega)\varphi^T(t,\omega)d\mu, \tag{9}$$

where $\varphi^T \in C^0([0,T];L^2(\Omega,\mathbb{R}^n))$ is the solution of the backward problem

$$\begin{cases} -\varphi_t^T + A^*(\omega)\varphi^T = C^*(\omega)(C(\omega)x^T - z) \quad t \in (0,T), \\ \varphi^T(T,\omega) = 0. \end{cases} \tag{10}$$

In what follows, we assume the following two conditions concerning the dynamics and the cost functional.

Hypothesis 1: For the pair (A, C) we assume the following condition. There exists a feedback operator $K_C \in C^0(\Omega, \mathcal{L}(\mathbb{R}^n))$ uniformly bounded with respect to ω such that, and there exist $\alpha > 0$ such that

$$\langle (A + K_C C)v, v\rangle_{L^2(\Omega,\mathbb{R}^n)} \geq \alpha \|v(\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n)}, \tag{11}$$

for all $v \in L^2(\Omega,\mathbb{R}^n)$.

Hypothesis 2: For the pair (A^*, B^*) we consider the next assumption. There exists $\kappa_1 \in \mathbb{R}$ and $\kappa_2 > 0$ such that

$$\langle A^*v, v\rangle_{L^2(\Omega,\mathbb{R}^n)} + \kappa_1 \left\| \int_\Omega B^*(\omega)v(\omega)d\mu \right\|^2_{\mathbb{R}^m} \geq \kappa_2 \|v(\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n)}, \tag{12}$$

for all $v \in L^2(\Omega,\mathbb{R}^n)$.

These assumptions are closely related to exponential stabilizability and exponential detectability, as mentioned in Grüne et al. (2019) for abstract differential equations. Under the previous assumptions we have the following "observability" estimates for $x^T(T)$ and $\varphi^T(0)$.

Lemma 3. Let us assume that **Hypothesis 1** holds. Then, there exists a constant $K > 0$ independent of $T > 0$ such that, for every $t \in [0,T]$

$$\|x^T(t,\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n)} \leq K\bigg(\int_0^T \Big[\|B(\cdot)u^T(t)\|^2_{L^2(\Omega,\mathbb{R}^n)} + \|C(\cdot)x^T(t,\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n)} \Big] dt + \|x_0\|^2_{\mathbb{R}^n} \bigg), \tag{13}$$

where (u^T, x^T) is the optimal pair given by Theorem 2.

Lemma 4. Let us assume that **Hypothesis 2** holds. Then, there exists a constant $K > 0$ such that for every $t \in [0,T]$

$$\|\varphi^T(t,\cdot)\|^2_{L^2(\Omega,\mathbb{R}^n)} \leq K \int_0^T \Big[\|C(\cdot)x^T(t,\cdot) - z\|^2_{L^2(\Omega,\mathbb{R}^n)} + \left\| \int_\Omega B^*(\omega)\varphi^T(t,\omega)d\mu \right\|^2_{\mathbb{R}^m} \Big] dt, \tag{14}$$

where φ^T is the solution of (10).

2.2 Stationary problem

We continue analyzing the stationary optimal control problem (3)–(4).

Under the **Hypothesis 2** we have the following existence and uniqueness for the optimal pair for problem (3)–(4).

Theorem 5. Assume that **Hypothesis 2** holds true. Then the problem (3)–(4) admits a unique optimal pair $(\overline{u}, \overline{x}) \in \mathbb{R}^m \times L^2(\Omega,\mathbb{R}^n)$, with $\overline{x}$ the optimal state associated to $\overline{u}$.

Now, we define the following set

$$D := \{u \in \mathbb{R}^m \ : \ B(\omega)u \in \mathrm{Ran}(A(\omega)), \text{ for each } \omega \in \Omega\}. \tag{15}$$

Theorem 6. Assume that **Hypothesis 2** holds true and let $(\overline{u}, \overline{x})$ be the unique solution of the optimal control (3)–(4). Then, there exists $\overline{\varphi} \in L^2(\Omega,\mathbb{R}^n)$ such that for a.e. $\omega \in \Omega$ we have

$$A^*(\omega)\overline{\varphi} = C^*(\omega)(C(\omega)\overline{x} - z), \tag{16}$$

and

$$\langle \overline{u}, v\rangle_{\mathbb{R}^m} + \langle \overline{\varphi}, Bv\rangle_{L^2(\Omega,\mathbb{R}^n)} = 0, \quad \forall v \in D. \tag{17}$$

3. MAIN RESULTS

In this section we state and prove the main results of this work. For this, let us recall the evolutionary and stationary optimality systems. Let (u^T, x^T) be the optimal pair of (7)–(8), and $(\overline{x}, \overline{u})$ the optimal pair of (3)–(4) (see Theorems 2 and 6). In addition, we have that there exist φ^T solution of (10), such that the optimal control u^T of (7) is given by

$$u^T(t) = -\int_\Omega B^*(\omega)\varphi^T(t,\omega)d\mu, \tag{18}$$

and $\overline{u}$ the optimal control associate to (3) satisfies

$$\langle \overline{u}, v\rangle_{\mathbb{R}^m} + \langle \overline{\varphi}, B(\omega)v\rangle_{L^2(\Omega,X)} = 0, \quad \forall v \in D, \tag{19}$$

where $\overline{\varphi}$ is the solution of (16). Besides, the following optimality systems hold:

$$\begin{cases} x_t^T + A(\omega)x^T = B(\omega)u^T \quad t \in (0,T), \\ -\varphi_t^T + A^*(\omega)\varphi^T = C^*(\omega)(C(\omega)x^T - z) \quad t \in (0,T), \\ x^T(0,\omega) = x_0, \quad \varphi^T(T,\omega) = 0, \end{cases} \tag{20}$$

and

$$\begin{cases} A(\omega)\overline{x} = B(\omega)\overline{u}, \\ A^*(\omega)\overline{\varphi} = C^*(\omega)(C(\omega)\overline{x} - z). \end{cases} \tag{21}$$

The first main result concerns the average convergence of the optimal pair (u^T, x^T) to the corresponding stationary ones $(\overline{u}, \overline{x})$, stated in the following theorem. The proof is based on the results contained in Porretta and Zuazua (2013).

Theorem 7. Let us assume that **Hypothesis 1 and 2** hold. Then,

$$\frac{1}{T}\int_0^T x^T(t,\cdot)dt \longrightarrow \overline{x}(\cdot) \quad \text{in } L^2(\Omega,\mathbb{R}^n),$$
$$\frac{1}{T}\int_0^T u^T(t)dt \longrightarrow \overline{u} \quad \text{in } \mathbb{R}^m,$$

as $T \to \infty$, where (u^T, x^T) is the optimal pair of (7)–(8), and $(\overline{x}, \overline{u})$ is the optimal pair of (3)–(4).

Corollary 8. Let us assume that **Hypothesis 1 and 2** hold. Then, there exists a unique $\overline{\varphi}$ solution of (16). In addition, the stationary optimal control $\overline{u}$ is given by

$$\overline{u} = -\int_\Omega B^* \overline{\varphi} d\mu.$$

Our second main result, which is the following theorem, shows the average exponential turnpike property. The proof is inspired by the results obtained in Grüne et al. (2019).

Theorem 9. Let us assume that **Hypothesis 1 and 2** hold. Let $\delta \geq 0$ be a real number. Let (u^T, x^T, φ^T) be the solution of (20) and $(\overline{u}, \overline{x}, \overline{\varphi})$ the corresponding stationary solution of (21). Then, there exists a positive constant $K = K(\delta) > 0$ (independent of T) such that

$$\left\|x^T(t,\cdot) - \overline{x}(\cdot)\right\|^2_{L^2(\Omega,\mathbb{R}^n)} + \left\|\varphi^T(t,\cdot) - \overline{\varphi}(\cdot)\right\|^2_{L^2(\Omega,\mathbb{R}^n)} \leq K(e^{-\delta(T-t)} + e^{-\delta t}), \quad (22)$$

for every $t \in (0,T)$. In particular, we obtain an averaged exponential turnpike as follows

$$\left\|\mathbb{E}(x^T - \overline{x})\right\|_{\mathbb{R}^n} + \left\|\mathbb{E}(\varphi^T - \overline{\varphi})\right\|_{\mathbb{R}^n} + \left\|u^T - \overline{u}\right\|_{\mathbb{R}^m} \leq K(e^{-\delta(T-t)} + e^{-\delta t}),$$

for every $t \in (0,T)$.

Remark 10. (1) It is immediately noted that in Theorems 7 and 9 one can also consider the case where $x_0, z \in L^2(\Omega,\mathbb{R}^n)$. The proof of both Theorems applies replacing the terms $\|x_0\|_{\mathbb{R}^n}, \|z\|_{\mathbb{R}^n}$ by $\|x_0(\cdot)\|_{L^2(\Omega,\mathbb{R}^n)}, \|z\|_{L^2(\Omega,\mathbb{R}^n)}$, respectively.

(2) Finally, it is interesting to note that our second main result, namely estimates (22) in Theorem 9, means that the turnpike holds for each parameter separately. This result is a strong consequence of our main assumptions Hypothesis 1 and 2. Also is interesting that this holds with optimal control, which is independent of random parameters. However, it captures, at the same time, all the information of the adjoint system (in an average sense).

4. NUMERICAL EXPERIMENTS

This section will perform numerical experiments to validate the average turnpike property. We focus our attention on the particular case $A(\omega) = \alpha(\omega)A$ and $B(\omega) = \beta(\omega)B$, with A and B constant matrices and α, β scalar random variables. Besides, the observability matrix will be independent of ω. In addition, we consider a discrete sample space Ω.

Let $\Omega = \mathbb{R}^+$ and β be a random variable with exponential distribution with parameter $\lambda = 7$ i.e. $\beta \sim exp(7)$ and $\alpha \sim Unif([1/2, 2])$. We consider the following optimal control problem

$$\min\left\{J^T(u) = \frac{1}{2}\left(\int_0^T \|u(t)\|^2_{\mathbb{R}^m} + \|Cx(t,\cdot) - z\|^2_{L^2(\Omega,\mathbb{R}^n)}\, dt\right)\right\},$$

subject to x solves the system

$$\begin{cases} x_t + \alpha(\omega)Ax = \beta(\omega)Bu & t \in (0,T), \\ x(0) = x_0. \end{cases}$$

where A, B and C satisfy the **Hypothesis 1 and 2**. The corresponding stationary optimal control problem is

$$\min\left\{J^s(u) = \frac{1}{2}\left(\|u\|^2_{\mathbb{R}^m} + \|Cx(\cdot) - z\|^2_{L^2(\Omega,\mathbb{R}^n)}\right)\right\},$$

subject to x solves the problem $\alpha(\omega)Ax = \beta(\omega)Bu$. We compute the optimal solutions (x^T, u^T) in time $T = 10$, and $(\overline{x}, \overline{u})$, by using the Gekko library on Python and considering seven realizations of the random variables β and α, which were generated using the numpy.random library on Python.

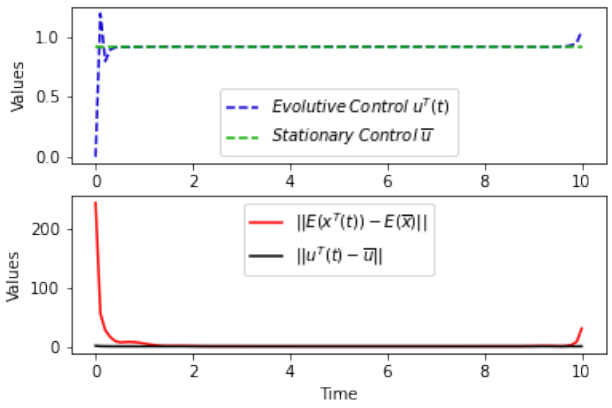


Fig. 1. Norm difference between evolutionary and stationary solutions and controls.

REFERENCES

Bergmann, P.G. (1946). Propagation of radiation in a medium with random inhomogeneities. *Phys. Rev.*, 70, 486–492. doi:10.1103/PhysRev.70.486.

Dorfman, R., Samuelson, P.A., and Solow, R.M. (1987). *Linear programming and economic analysis.* Dover Publications, Inc., New York. Reprint of the 1958 edition.

Grüne, L., Schaller, M., and Schiela, A. (2019). Sensitivity analysis of optimal control for a class of parabolic PDEs motivated by model predictive control. *SIAM J. Control Optim.*, 57(4), 2753–2774. doi:10.1137/18M1223083.

Lohéac, J. and Zuazua, E. (2016). From averaged to simultaneous controllability. *Ann. Fac. Sci. Toulouse Math. (6)*, 25(4), 785–828. doi:10.5802/afst.1511.

Porretta, A. and Zuazua, E. (2013). Long time versus steady state optimal control. *SIAM J. Control Optim.*, 51(6), 4242–4273. doi:10.1137/130907239.

Soong, T.T. (1973). *Random differential equations in science and engineering.* Mathematics in Science and Engineering, Vol. 103. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London.

Soong, T.T. (1981). *Probabilistic modeling and analysis in science and engineering.* John Wiley & Sons, Inc., New York.

Sun, J., Wang, H., and Yong, J. (2022). Turnpike properties for stochastic linear-quadratic optimal control problems. *arXiv preprint arXiv:2202.12699.*

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

An Overview of NASA's Learn-to-Fly Concept ⋆

Steven Snyder *

* *NASA Langley Research Center, Hampton, VA 23681 USA*
(e-mail: steven.m.snyder@nasa.gov).

Abstract: Learn-to-Fly is a framework for incorporating learning methods into the development cycle of new aircraft. This paper provides an overview of the concept and summarizes some results from previous test campaigns. It will also discuss some perceived benefits of Learn-to-Fly and improvements to the procedure for more practical and widespread use. Ongoing efforts are needed to continue to develop the necessary underlying technologies for integration into this framework.

Keywords: Learning methods, Autonomous control, Aircraft system identification, Real time.

1. INTRODUCTION

NASA's Learn-to-Fly (L2F) concept seeks to apply learning methods to the modeling and control of aircraft. The current paradigm for aerodynamic modeling and control law design of a new vehicle relies on significant input from ground-based methods, such as computational fluid dynamics (CFD) simulations and wind tunnel testing, to develop a simulation model that is then used for the development of the control law. With the control law designed based on the simulation model, flight tests are performed, and it is often determined from the flight data that there are areas requiring additional data for aerodynamic model improvement. Additional flight test sorties are planned until sufficient data has been captured and the control laws are improved to desired levels of performance and robustness. This iterative, sequential process requires many different subject matter experts to be engaged throughout the development of the aerodynamic model and control law. Conversely, the L2F concept is built upon real-time modeling, real-time guidance, and learning control that work together to identify the aerodynamic model and design the control law for the vehicle with minimal human input. By developing the aerodynamic model based on flight data, the control system is designed with actual flight dynamics responses rather than analogous results, removing the need for corrections due to Reynolds number (if flights are full-scale), wind tunnel blockage, boundary-layer turbulence, etc.

Recent advances in aerodynamic modeling allow such learning processes to operate onboard an aircraft in real time with commercially available computing hardware, e.g., see Morelli (2018, 2020). This allows the flight data content to be assessed in near real time while performing system identification maneuvers, providing a means of determining when a valid aerodynamic model has been generated. If an aerodynamic model of a vehicle can be determined in flight, then there is the potential for automatically modifying control laws, allowing the aircraft to autonomously "learn to fly" and improve its performance with additional flight time (experience). See Snyder et al. (2018), Snyder (2020), Weinstein et al. (2018), and Grauer (2018) for examples.

One would be remiss not to acknowledge the advances in ground testing capabilities. Over time, CFD simulations have become more accurate and cost efficient to perform, and CFD methods are expected to continue improving (see Cary et al. (2021)). Murphy and Brandon (2017) have applied modern design of experiment methods to improve wind tunnel test designs. Additional research has been performed on automating some of these test procedures to reduce the amount of engineering judgment and human input required, as described by Murphy et al. (2020, 2021). Given this, it seems likely that the future may hold a combination of automated ground testing and automated inflight modeling and control updates, like those in L2F.

The remainder of this paper includes a brief overview of the technologies within the current L2F framework in Section 2 along with some highlights of early L2F flight test results in Section 3. Finally, Section 4 contains a discussion of some open areas of research that should be filled to improve implementations of the L2F concept.

2. COMPONENTS OF THE LEARN-TO-FLY FRAMEWORK

As mentioned above, L2F contains three primary components: 1) real-time modeling, 2) real-time guidance, and 3) learning control. A block diagram showing the connections between these components and some of the data flow is depicted in Figure 1. Each of these pieces is described briefly below with references for more details on the implementation used in flight testing. Note that different algorithms that perform these functions could be swapped in and out of this framework.

Real-time Modeling
The goal of the real-time modeling module is to iden-

⋆ Current funding for Learn-to-Fly technologies is provided by NASA under the Transformational Tools and Technologies (TTT) project.

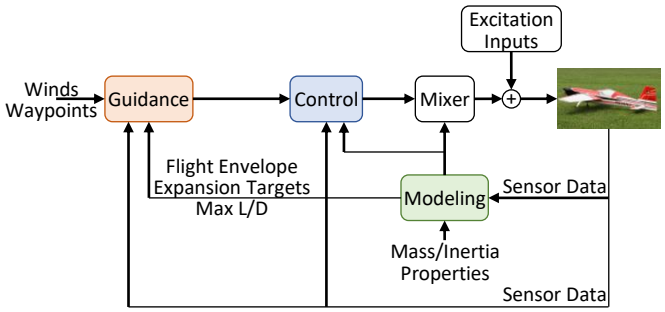


Fig. 1. Block diagram of the Learn-to-Fly architecture.

tify a six-degree-of-freedom mathematical model of the aerodynamic forces and moments acting on the vehicle. It is assumed that the rigid-body kinematic relations are known. The measured states and control surface positions are used as the explanatory variables for the aerodynamic forces and moments. Non-dimensional forces and moments depend not only on the measured aircraft states, but also on the vehicle geometry and mass properties. Since these cannot be measured directly during flight, the geometry and mass properties must be known a priori. Two factors play a role in the creation of an accurate global aerodynamics model during flight: safe and efficient flight maneuvers to sufficiently excite the targeted aircraft dynamics, and a recursive system identification algorithm that can be implemented in real time.

The maneuvers used for excitation were created by injecting orthogonal phase-optimized multi-sine inputs into the control surface commands during portions of the flight tests, as discussed by Morelli (2018, 2020). These inputs are referred to as programmed test inputs (PTIs). The phase optimization indirectly minimizes the peak amplitude, and they only result in a small net perturbation to the aircraft dynamics because the inputs have zero mean. Because these signals are orthogonal, the effect of the inputs can be determined despite simultaneous actuation. This provides a rich amount of excitation over a small window of time.

The recursive system identification algorithm that was implemented is described in detail by Morelli (2018). The approach is based on a candidate pool of multivariate orthogonal functions, referred to as regressors. These functions can be of arbitrary complexity and are selected a priori by the designer. Theoretically, the pool of candidates could be arbitrarily large, yet in practice, the size is limited by onboard computing capabilities. A recursive QR decomposition is used to recursively orthogonalize the regressors in order to isolate and quantify their explanatory capabilities of the dynamics. A least-squares estimator then calculates coefficients for each orthogonalized regressor by minimizing the sum of squared differences between the computed non-dimensional forces and moments, and those predicted by the model. Measures of model fit quality and a penalty for model complexity are then balanced to determine the model structure, i.e., the orthogonalized regressors that will be included within the model. Finally, the model is transformed back into physical quantities by reversing the orthogonalization procedure. Applying this multivariate orthogonal function modeling to each of the rigid-body degrees of freedom will result in the desired six-degree-of-freedom mathematical model. Uncertainty bounds corresponding to each model parameter estimate can also be computed from the generated statistics. A more thorough discussion of the time-domain modeling procedure is given by Morelli (2018). Morelli (2020) provides updates to the modeling procedure in the frequency domain.

Real-time Guidance

Several functions were overseen by the real-time guidance algorithm, including waypoint navigation, energy management for autonomous landing, and limited flight envelope protection. Some degree of autonomous envelope expansion was also provided by the guidance, which systematically adjusted the nominal angle of attack and sideslip commands so that the PTIs could excite the vehicle dynamics in different parts of the flight envelope. By monitoring the real-time global model for instabilities, the guidance algorithm could limit the envelope expansion in order to preserve system safety and keep the vehicle within the stable portion of the flight envelope. Navigation was performed by computing the ground track to the desired waypoint. After flying through a given acceptance radius of the waypoint, the guidance would advance to the next waypoint and compute the ground track relative to it.

When flown on a glider-type aircraft, an altitude trigger would switch the guidance commands over to a landing mode. In this mode, the longitudinal guidance command would switch from angle of attack to flight path angle. The optimal angle, in terms of maximizing the ratio of lift to drag, would be calculated from the real-time aerodynamic model and sent to the control law.

For a powered vehicle flyable by a radio control (R/C) pilot, takeoffs and landings can be performed by the pilot. An altitude and speed profile can be provided along with the waypoint tracking, and the guidance algorithm can then compute a desired flight path angle and speed to demand from the control law. Additionally, with the capability of piloted flight, the real-time modeling can be performed either under manual control or autonomous control. More details on the guidance algorithm and capabilities are given by Foster (2018).

Learning Control

The main objective of the vehicle's control law is to robustly stabilize the vehicle and track the provided guidance commands. The controller must handle disturbances and uncertainties within the system. It uses a traditional adaptive control architecture, complete with a reference model, to react to any immediate disturbances or unknowns. However, as patterns emerge within these unknowns, the real-time modeling results capture these patterns, making those unknowns known. The baseline controller and reference model are modified based on the real-time modeling results, and the system no longer needs to react to those specific disturbances since it can now predict them. The distinction here is between modifying behavior based on a reaction to something that has already happened (adaptation) and modifying behavior based on a prediction of what will happen (learning). Past experiences are required to generate an appropriate prediction.

Some amount of perturbation from the flight condition is required for the real-time modeling to learn the aerodynamics. This is provided by the multi-sine inputs described

above, yet the excitation provided by these inputs goes contrary to the controller's goal of stabilization and are seen by the controller as a disturbance. Simulation studies during the L2F development have shown that rejecting too much disturbance reduced the real-time aerodynamic modeling's ability to identify a model. These conflicting goals of identification and stabilization need to be balanced throughout the L2F process.

The structure of the controller used contains two elements, a baseline controller and an adaptive augmentation. For the baseline controller, a nonlinear dynamic inversion (NDI) control law was designed based on the model learned in real time. NDI essentially seeks to cancel any undesirable dynamics within the determined mathematical model. The desired dynamics after inversion were chosen such that the natural frequency of the vehicle is preserved and sufficient damping is provided for tracking the guidance commands. To augment the baseline controller, a model-based adaptive law compares the measured response to the predicted response of the desired dynamics, adjusting the control signal to reduce any discrepancies. Additional details and results on the control laws are given by Snyder et al. (2018).

Due to the periodic updates of the learned model, the reference model (desired dynamics) exhibits a switching behavior. Switching makes the theoretical analysis of the model-based adaptive controller more challenging than for simple linear time-invariant systems, but an initial analysis framework for an $\mathcal{L}_1$ adaptive controller for this class of systems was developed by Snyder et al. (2022).

3. FLIGHT TESTS

The flight tests discussed here were performed on a 40%-scale, conventional, powered aircraft. The vehicle, known as E1, was modified to split the flaperons into conventional flaps and ailerons and to split the elevator into independent left and right surfaces. These extra control surfaces enabled the use of a hidden feedback loop to modify the apparent dynamics of the vehicle. This feedback loop could, for example, command the left elevator to destabilize the pitch channel based on the angle of attack, unbeknownst to the modeling, guidance, or control modules. The destabilization could be used to create a static instability in pitch or, by feeding back the roll rate to the flaps, create unstable roll damping. Thus, the L2F method could be tested on an effectively unstable vehicle. Figure 2 shows how the surfaces were allocated between stability degradation and control.

Of course, for piloted takeoffs and landings, it is undesirable to have an unstable vehicle. Therefore, the R/C pilot was provided the ability to engage or disengage this destabilizing feedback. Three modes were available to the pilot. The first mode bypassed the L2F computer and allowed the pilot to fly with conventional R/C avionics. The second mode passed the pilot commands through the L2F computer, allowing the pilot commands to be augmented, such as with the destabilizing feedback. The final mode was for fully autonomous flight.

Select Test Results

In the tests described here, the pilot took off in the bypass

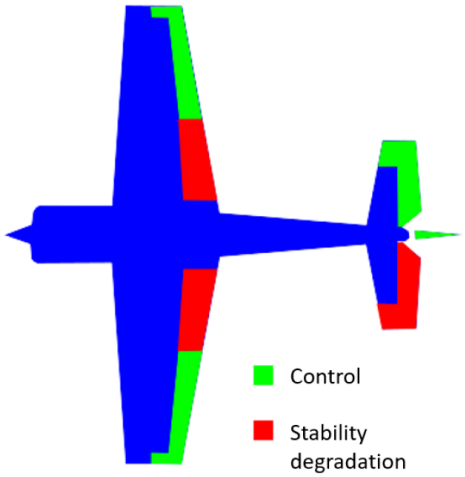


Fig. 2. Schematic of how control surfaces were allocated between control and stability degradation.

mode. Upon reaching the specified test altitude, he then either engaged the autonomous mode, which enabled both the modeling PTIs and the destabilization feedback loop along with the L2F software, or he engaged the pilot pass through mode, which injected the PTIs on top of the pilot inputs and also enabled the destabilizing feedback. The vehicle was flown twice with the pitch destabilization, once piloted and once autonomously.

Figure 3 overlays the responses of the autonomous L2F system and the pilot. A post-flight analysis of the data determined that the static margin of the vehicle was approximately -16.4%, which is a significant instability. This can explain why the R/C pilot had difficulty getting the pitch angle under control, even with large inputs. The pilot traces within the figure are short because the pilot felt uncomfortable attempting to fly the destabilized vehicle through a turn that was required for range safety purposes. As such, he reverted to the safety mode to regain control of the vehicle. With the autonomous L2F system running, some initial pitch oscillations occur. The pitch angle θ of the vehicle can be seen in the first plot, along with the elevator deflection (input signal) δ_e in the second plot as the controller tries to stabilize the aircraft. While the real-time modeling procedure does not rely on any sort of initial guess of the vehicle dynamics, the control laws do. In each case, the initial guess provided to the controller was of a stable aircraft, which can be seen by the initial negative value of the model's pitch coefficient due to angle of attack, C_{m_α}, in the third plot. Thus, the initial pitch oscillations are not surprising.

After about 1 s, the autonomous system begins learning the vehicle dynamics and updating the model parameters (highlighted area). Remember that learning requires experience. This is why one cannot expect the model to update instantaneously. It must first see the pattern in order to learn it. The principal frequency of the pitch dynamics of this vehicle is roughly 1 Hz. Notice that it takes about two cycles (approximately 2 seconds) for the real-time modeling to learn what it is seeing and converge appropriately. Based on the updated model, the gains of the control law, as represented by K_α in the fourth plot, are also updated.

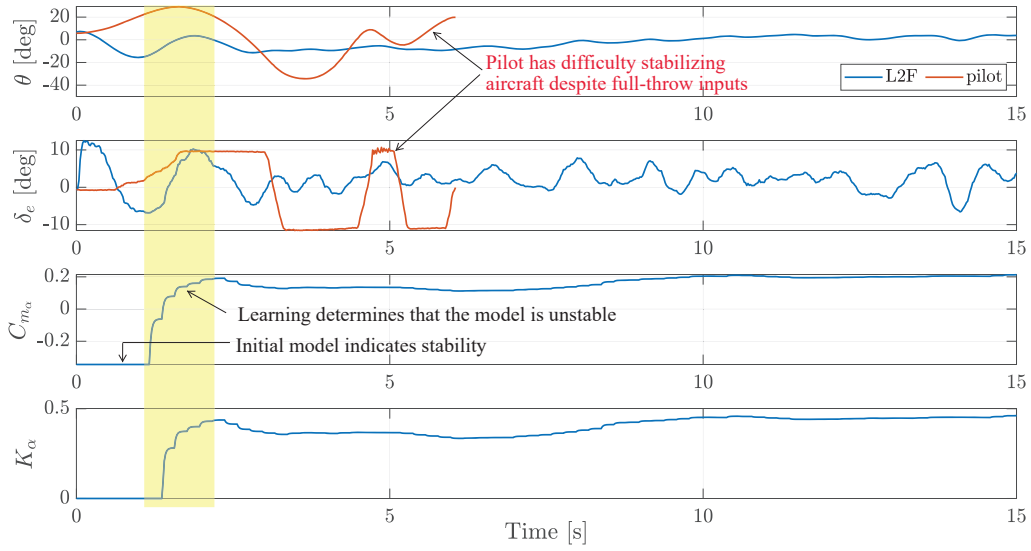


Fig. 3. Overlay of two flights with pitch destabilization, one with the pilot and one with the autonomous L2F system.

4. DISCUSSION

Learn-to-Fly is an interesting application area for learning-based methods in control and one that could benefit the aviation industry. However, many of the improvements needed to push L2F forward are the same developments generally needed for higher level autonomous agents.

Imagine cooperative agents whose objectives compete, at least temporarily. How can those conflicts be resolved so that the team can thrive, especially in cases when the agents are not just cooperative but also codependent? In the L2F scenario, the modeling and control modules are exactly that—codependent. Without model information, control of the vehicle could easily be lost, and if the vehicle is out of control—whether tumbling or possibly crashing—meaningful models cannot be produced.

Many systems that contain some sort of learning element will always have its learning engaged or have it be controlled by a human user. If the system response is consistent with the autonomous agent's expectation, then the learning process need not be active, freeing up resources. However, some level of monitoring would likely be prudent in case the expectation and measured response should, at some future point, diverge. How can resources be balanced between learning and doing? For a system like L2F, it is worth noting that different levels of learning could be applied, such as self learning—which has been the focus thus far—and external learning, which is learning about the environment or other agents within it. Balance would need to be struck between these tasks as well.

ACKNOWLEDGEMENTS

The hard work and dedication of the Learn-to-Fly team are gratefully acknowledged.

REFERENCES

Cary, A., Chawner, J., Duque, E., Gropp, W., Kleb, W., Kolonay, R., Nielsen, E., and Smith, B. (2021). CFD vision 2030 road map: Progress and perspectives. In *AIAA Aviation Forum*. doi:10.2514/6.2021-2726.

Foster, J. (2018). Autonomous guidance algorithms for NASA learn-to-fly technology development. In *AIAA Aviation Forum*. Atlanta, GA. doi:10.2514/6.2018-3310.

Grauer, J. (2018). A learn-to-fly approach for adaptively tuning flight control systems. In *AIAA Aviation Forum*. Atlanta, GA. doi:10.2514/6.2018-3312.

Morelli, E. (2018). Practical aspects of real-time modeling for the learn-to-fly concept. In *AIAA Aviation Forum*. Atlanta, GA. doi:10.2514/6.2018-3309.

Morelli, E. (2020). Autonomous real-time global aerodynamic modeling in the frequency domain. In *AIAA SciTech Forum*. Orlando, FL. doi:10.2514/6.2020-0761.

Murphy, P. and Brandon, J. (2017). Efficient testing combining design of experiment and learn-to-fly strategies. In *AIAA SciTech Forum*. Grapevine, TX. doi:10.2514/6.2017-0696.

Murphy, P., Buning, P., and Simmons, B. (2021). Rapid aero modeling for urban air mobility aircraft in computational experiments. In *AIAA SciTech Forum*. doi:10.2514/6.2021-1002.

Murphy, P., Hatke, D., Aubuchon, V., Weinstein, R., and Busan, R. (2020). Preliminary steps in developing rapid aero modeling technology. In *AIAA SciTech Forum*. Orlando, FL. doi:10.2514/6.2020-0764.

Snyder, S. (2020). Autopilot design with learn-to-fly. In *AIAA SciTech Forum*. Orlando, FL. doi:10.2514/6.2020-0763.

Snyder, S., Bacon, B., Morelli, E., Frost, S., Teubert, C., and Okolo, W. (2018). Online control design for learn-to-fly. In *AIAA Aviation Forum*. Atlanta, GA. doi:10.2514/6.2018-331.

Snyder, S., Zhao, P., and Hovakimyan, N. (2022). $\mathcal{L}_1$ adaptive control with switch reference models: Application to learn-to-fly. *Journal of Guidance, Control, and Dynamics*. Under review.

Weinstein, R., Hubbard, J., and Cunningham, M. (2018). Fuzzy modeling and parallel distributed compensation for aircraft flight control from simulated flight data. In *AIAA Aviation Forum*. Atlanta, GA. doi:10.2514/6.2018-3313.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On rank metric convolutional codes and concatenated codes ⋆

Diego Napp * **Raquel Pinto** ** **Carlos Vela** **

* *Department of Mathematics, University of Alicante, Alicante, Spain (e-mail: diego.napp at ua.es).*
** *Department of Mathematics, University of Aveiro, 3810-197 Aveiro, Portugal (e-mail: carlos.vela, raquel at ua.pt)*

Abstract: In the recent history of the theory of network coding the multi-shot network coding has been prove as a good alternative for the classical one-shot network theory which is managed by using block codes. To perform communications in this multi-shot context we have, among others, rank-metric convolutional codes and concatenated codes (using a convolutional code as an outer code and a rank-metric code as inner code). In this work we analyse their performance over the rank deficiency channel (described by Gilbert-Elliot channel model) in terms of the correction capabilities and the complexity of the two decoding schemes.

Keywords: concatenated codes, rank-metric codes, convolutional codes, rank-deficiency channel, complexity

1. INTRODUCTION

Since its raising at the beginning of 2000's, network coding has been a research topic that has attracted significant interest in many areas, including electrical engineering, computer science and applied mathematics. Network coding theory provides a pragmatic instrument to disseminate information (packets) over networks where there may be many information sources and possibly many receivers. From a mathematical point of view, these packets can be modelled by columns of matrices over a finite field $\mathbb{F}_q$ and during the transmission, these columns are linearly combined at each node of the network. To achieve reliable communication over this channel, *rank-metric codes* are typically employed.

Most of the literature deals with the situation in which the network is used only once to propagate the information. Such scenario is referred to as *one-shot* network coding, as the encoding and transmission is performed over one use (shot) of the network. If one needs to transmit more data (packets), then these packets are again encoded and transmitted in the following instant, independently on the previous transmissions. However, one can improve the error-correction capability of the code in the scenario where we need to use the network several times (*multi-shot*) by creating correlation among the transmitted data in different shots. This new approach has recently attracted much attention due to possible interesting applications, e.g., in streaming communications Mahmood et al. (2015). Nevertheless, network coding techniques for streaming are fundamentally different from the classical ones. To be optimised they must operate under low-latency, sequential encoding and decoding constraints, and as such they must inherently have a convolutional structure. That is the reason why most of the proposed schemes for this scenario employ convolutional codes in different ways Wachter-Zeh et al. (2015); Napp et al. (2017a); Mahmood et al. (2015); Almeida et al. (2020).

In this work we present a comparison of two different and important schemes for multi-shot network coding: rank metric convolutional codes and concatenated codes (concatenation of a convolutional code and a rank metric code) Napp et al. (2018). In particular, we compare their performance over a rank-deficiency channel focusing on the correction capabilities and the complexity of the encoding and decoding procedures of each proposal. For practical reasons, in our comparison we limit the field size. Since the construction of optimal rank-metric codes for this channel exists Mahmood et al. (2015) only for large finite fields, such codes cannot be used in this context. The concatenated codes, however, can be constructed optimally. In Section 2 we present the necessary background about convolutional codes, rank-metric convolutional codes and concatenated codes and the nature of the used channel. In Section 3 we make a comparison of the performance of both codes over this kind of channel. Later, in Section 4 we compare the complexity in the decoding process under the circumstances established in the previous section. Finally, in Section 5 we make some review of this work and present future possible works.

2. PRELIMINARIES

2.1 Convolutional codes

Let $\mathbb{F}_q$ be a finite field and $\mathbb{F}_q[D]$ the ring of polynomials with coefficients in $\mathbb{F}_q$. A *convolutional code* $\mathcal{C}$

⋆ This work has been partially supported by the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UIDB/04106/2020. The second author was also supported by the Ministerio de Ciencia e Innovación under the grant with ref. PID2019-108668GB-I00.

of rate k/n is a rank k $\mathbb{F}_q[D]$-submodule of $\mathbb{F}_q[D]^n$. If $G(D) \in \mathbb{F}_q[D]^{k\times n}$ is a full row rank matrix such that

$$\mathcal{C} = \mathrm{Im}_{\mathbb{F}_q[D]} G(D) = \left\{ u(D)G(D) : u(D) \in \mathbb{F}_q[D]^k \right\},$$

then $G(D)$ is called an encoder of $\mathcal{C}$.

Any other encoder $\tilde{G}$ of $\mathcal{C}$ differs from $G(D)$ by a unimodular matrix $U(D) \in \mathbb{F}_q[D]^{k\times k}$. *i.e.*, $\tilde{G}(D) = U(D)G(D)$. Therefore, we can consider $G(D)$ to be minimal, *i.e.*, in row reduced form Johannesson and Zigangirov (1999). In this case, the sum of the row degrees of $G(D)$ attains its minimum among all the encoders. This value is usually denoted by δ and called the degree of $\mathcal{C}$. A convolutional code with rate k/n and degree δ is called an (n, k, δ) convolutional codes, McEliece (1998). The largest row degree over one, and therefore all, reduced encoders of $\mathcal{C}$ is called the memory of $\mathcal{C}$ and denoted by μ. If the memory is considered instead of the degree, a convolutional code with rate k/n and memory μ is referred to as an (n, k, μ) convolutional code Mahmood et al. (2015).

An important distance measure for convolutional codes $\mathcal{C}$ is its free distance which is defined as

$$d_{free}(\mathcal{C}) = \min_{v(D)\in\mathcal{C}, v(D)\neq 0} wt(v(D)),$$

where $wt(v(D))$ is the Hamming weight of a polynomial vector $v(D) = \sum_{i\in\mathbb{N}} v_i D^i \in \mathbb{F}_q[D]^n$, defined as

$$wt(v(D)) = \sum_{i\in\mathbb{N}} wt(v_i),$$

being $wt(v_i)$ the number of the nonzero components of v_i. Another important distance measure is the *j-th column distance*,

$$d_H^j(\mathcal{C}) = \min\left\{ wt(v(D)|_{[0,j]}) \,|\, v(D) \in \mathcal{C}, v_0 \neq 0 \right\}$$

where $v(D) = \sum_{i\in\mathbb{N}} v_i D^i$ and $v(D)|_{[0,j]} = \sum_{i=0}^{j} v_i D^i$. This measure is also upper-bounded in Johannesson and Zigangirov (1999):

$$d_H^j(\mathcal{C}) \leq (n-k)(j+1)+1$$

for $j \leq L$ where $L = \lfloor \delta/k \rfloor + \lfloor \delta/(n-k) \rfloor$. The convolutional code achieving the bound for all $j \in \{0, \ldots, L\}$ is called *maximum distance profile* (MDP) convolutional code Gluesing-Luerssen et al. (2006).

2.2 Rank-metric codes

Let $A, B \in \mathbb{F}_q^{n\times m}$. The rank distance between two matrices is

$$d_{rank}(A, B) = rank(A - B).$$

This defines a distance, called *rank distance*, and rank-metric codes are subsets of $\mathcal{C} \subseteq \mathbb{F}_q^{n\times m}$ equipped with the rank distance. Rank metric codes in $\mathbb{F}_q^{n\times m}$ are usually constructed as block codes of length n over the extension field $\mathbb{F}_{q^m}$ as in Kötter and Kschischang (2008). For a given basis of $\mathbb{F}_{q^m}$ viewed as an m vector space over $\mathbb{F}_q$, any element of $\mathbb{F}_{q^m}$ can be seen as a vector in $\mathbb{F}_q^m$. In this paper we will follow this approach and consider rank-metric codes as linear codes $[n, k]$ over $\mathbb{F}_{q^m}$. For the sake of simplicity we assume that $m \leq n$. In this case, linear rank metric codes of length n and dimension k over $\mathbb{F}_{q^m}$ must satisfy the following Singleton-type bound:

$$d_{rank}(\mathcal{C}) \leq n - k + 1.$$

A code that achieves this bound is called *Maximum Rank Distance* (MRD). Gabidulin codes are a well-known class of MRD codes as showed in Gabidulin (1985).

2.3 Rank-metric convolutional codes

Rank metric convolutional codes over $\mathbb{F}_{q^m}$ were first introduce in Wachter-Zeh et al. (2015) for unitmemory codes and for unrestricted memory in Mahmood et al. (2015); Almeida et al. (2020). These are convolutional codes defined over an extension field $\mathbb{F}_{q^m}$ and equipped with a rank-type metric, and as such, are referred to as (n, k, δ)-rank metric convolutional codes (over $\mathbb{F}_{q^m}$) if have length n, dimension k and degree δ. Later, a wider definition of convolutional codes over $\mathbb{F}_q$ (instead of over $\mathbb{F}_{q^m}$) was proposed in Napp et al. (2017b).

In Mahmood et al. (2015) a new column rank-base distance is considered. Let $\mathcal{C}$ be a (n, k, μ) be a convolutional code over $\mathbb{F}_{q^M}$. The *j-th column rank distance* of $\mathcal{C}$ is:

$$d_{SR}^j(\mathcal{C}) = \min_{x_{[0,j]}\in\mathcal{C}} \sum_{t=0}^{j} rank(\phi_n(x_{[0,j]}))$$

where $\phi_n : \mathbb{F}_{q^M}^n \to \mathbb{F}_q^{n\times M}$ is the bijective mapping which allows to use the rank based metric instead of the Hamming metric. This column distance is upper-bounded by:

$$d_{SR}^j(\mathcal{C}) \leq (n-k)(j+1)+1.$$

The codes which achieves this bound are named *Maximum Sum Rank codes* (MSR). These codes are showed to exists and a construction is given Mahmood et al. (2015). The larger the column distance is, the better is the correction capability within an interval of time. Hence, rank metric convolutional codes with optimal column distance profile are ideal for fast decoding, *i.e.,* for streaming application with low delay constrains. For these reasons we shall consider in this work MSR convolutional codes for our analysis.

2.4 Concatenated codes

In this section we introduce a completely different class of codes in the context of multi-shot network coding. Such a scheme comprises the concatenation of a Hamming metric convolutional code as an "outer code" and a rank metric block code as an "inner code". These codes are described by the concatenation scheme proposed in Napp et al. (2018) as follows: Let $\mathcal{C}_I$ be a linear (n_I, k_I) rank metric code over $\mathbb{F}_{q^m}$ with (rank) distance $d_{rank}(\mathcal{C}_I)$ and generator matrix $G_I \in \mathbb{F}_{q^m}^{k_I\times n_I}$. Let $\mathcal{C}_O$ be an (n_O, k_O, δ) convolutional code over the field $\mathbb{F}_{q^{mk_I}}$ with Free distance $d_H(\mathcal{C}_O)$, column distance $d_H^j(\mathcal{C}_O)$ and a generator matrix $G_O(D) \in \mathbb{F}_{q^{mk_I}}[D]^{k_O\times n_O}$.

The information (row) vector $u(D) = u_0 + u_1 D + u_2 D^2 + \cdots \in \mathbb{F}_{q^{mk_I}}[D]^{k_O}$ is encoded through $G_O(D)$ to generate

$$v(D) = v_0 + v_1 D + v_2 D^2 + \cdots = u(D)G_O(D) \in \mathcal{C}_O.$$

We write $v_i = (v_i^0, v_i^1, \ldots, v_i^{n_O-1}), v_i^j \in \mathbb{F}_{q^{mk_I}}$. Now, we identify $v_i^j \in \mathbb{F}_{q^{mk_I}}$ with a vector $\nu_i^j \in \mathbb{F}_{q^m}^{k_I}$ (for a given

basis of $\mathbb{F}_{q^{mk_I}}$) and write $\nu_i = (\nu_i^0, \nu_i^1, \ldots, \nu_i^{n_O-1}) \in (\mathbb{F}_{q^m}^{k_I})^{n_O}$ and therefore

$$\nu(D) = \nu_0 + \nu_1 D + \nu_2 D^2 + \cdots \in \mathbb{F}_{q^m}^{k_I}[D]^{n_O}.$$

Finally, the codewords $x(D)$ of the concatenated code $\mathcal{C}$ are obtained through the matrix $G_I \in \mathbb{F}_{q^m}^{k_I \times n_I}$ in the following way:

$$x_i^j = \nu_i^j G_I \in \mathbb{F}_{q^m}^{n_I},$$
$$x_i = (x_i^0, x_i^1, \ldots, x_i^{n_O-1}) \in (\mathbb{F}_{q^m}^{k_I})^{n_O},$$
$$x(D) = x_0 + x_1 D + x_2 D^2 + \cdots \in \mathcal{C} \subset \mathbb{F}_{q^m}^{k_I}[D]^{n_O}.$$

The sum rank metric and the j-th column sum rank distances are bounded by (see Napp et al. (2018)):

$$d_{SR}(\mathcal{C}) \leq (n_O n_I - k_O k_I)(\left\lfloor \frac{\delta}{k_O} \right\rfloor + 1) + \delta k_I + 1,$$

$$d_{SR}^j(\mathcal{C}) \leq (n_O n_I - k_O k_I)(j+1) + 1,$$

respectively.

2.5 Network model

In multi-shot network coding the information (packets) sent in the different uses of the network (shots) are correlated to improve the correction capability of the codes. The natural tool to use in this network are the convolutional codes which take into account this delay in the transference of the information. The network channel considered here is the *deficiency channel* which is a simplification of more general network channel and can be seem as the analogue of the erasure channel in the context of networks, see Mahmood et al. (2015) for more details. In this channel, at each shot the destination node observes $y_t = x_t A_t$, where $A_t \in \mathbb{F}_q^{n \times n}$ is the channel matrix at time t, and is known to the receiver, as explained in Ho et al. (2006). Communication over a window $[t, t+W-1]$ of W shots is described using $y_{[t,t+W-1]} = x_{[t,t+W-1]} A_{[t,t+W-1]}$, where $A_{[t,t+W-1]} = diag(A_t, \ldots, A_{t+W-1})$ is a block diagonal channel matrix as described in Mahmood et al. (2015). Let $\rho_t \triangleq rank(A_t)$ denote the rank of A_t, for all $t \geq 0$, we have that $\sum_{i=t}^{t+W-1} \rho_i = rank(A_{[t,t+W-1]})$. If during the circulation of the information, some of the intermediate nodes fails, the transmission will continue to work without including it in the linear combinations of the packets If all links are functional in the shot at any time t, then $\rho_t = n$, but failing links may result in a rank-deficient channel matrix at that time.

In the context of rank deficiency channels, we shall consider channels that satisfy certain conditions within an interval of time (window), see Mahmood et al. (2015). These are called *Rank-Deficient Sliding Window Networks*, denoted by $\mathcal{CH}(S, W)$ and have the property that in any sliding window of length W, the rank of the block diagonal channel decreases by no more than S, i.e., $\sum_{i=t}^{t+W-1} \rho_i \geq nW - S$ for each $t \geq 0$. We will say that a linear code $\mathcal{C}$ over $\mathbb{F}_{q^M}$ is defined as *feasible* for the channel $\mathcal{CH}(S, W)$ if the encoding and decoding functions for the code are capable of perfectly recovering every source packet transmitted over it with delay T, *i.e.*, to achieve the information of the packet x_t received at moment t by performing the necessary operations with at most the next T received packets, that is $x_t, \ldots, x_{t+T-1}$.

3. PERFORMANCE OF THE CODES

3.1 Discussion

In this subsection we will compare the correction capabilities of the two proposed codes over a rank-deficient sliding window channels. First, we will see under which constrains the cited codes are feasible for the channel $\mathcal{CH}(S, W)$. Secondly, we will compare their bounds on the j-th column distances and discuss the capabilities of them.

In Mahmood et al. (2015), a construction for the MSR convolutional codes is proposed. It is said that these codes are feasible for a rank-deficiency sliding window channel $\mathcal{CH}(S, W)$ with delay $T \geq W$ under the assumption $S < d_{SR}^{W-1}(\mathcal{C})$, where $\mathcal{C}$ is the MSR convolutional code. In the case in which $T < W$ it is enough to consider $S < d_{SR}^{T}(\mathcal{C})$. These codes guarantee the recover of the information under the worst channel conditions for a fixed delay and rate, i.e., they identify the largest rank deficiency S for which a code with a given rate is feasible.

On the other hand, we consider the concatenated codes constructed in Napp et al. (2018). Let $\mathcal{C}$, $\mathcal{C}_O$ and $\mathcal{C}_I$ be the concatenated code, the convolutional code and the rank-metric code, respectively as described above. The concatenated codes are also feasible for the rank-deficiency sliding window channel $\mathcal{CH}(S, W)$ with delay $T' \geq W$ if $S < d_H^{T'}(\mathcal{C}_O) d_{rank}(\mathcal{C}_I)$ where $T' = \left\lfloor \frac{W-1}{n_I} \right\rfloor$ due to the construction of the code. When $T < W$ it is enough to consider $S < d_H^{T'}(\mathcal{C}_O) d_{rank}(\mathcal{C}_I)$ where $T' = \left\lfloor \frac{T}{n_I} \right\rfloor$. The next result, establish which of these two families of codes have better distance bounds under the same conditions:

Theorem 1. Let $\mathcal{C}_{MSR}$ be a MSR convolutional code and C_{Conc}, $\mathcal{C}_O$ and C_I be a concatenated, convolutional and rank metric codes, respectively. Over a rank-deficiency sliding window channel $\mathcal{CH}(S, W)$, with fixed rate $k/n = k_O k_I / n_O n_I$ and delay T, then

$$d_H^{j'}(\mathcal{C}_O) d_{rank}(\mathcal{C}_I) < d_{SR}^{j-1}(\mathcal{C}_{MSR})$$

where $j' = \left\lfloor \frac{j-1}{n_I} \right\rfloor$ with $1 \leq j \leq T$.

Note that in the theorem above, we compare the corresponding column distance after receiving the same amount of shots. Since the bound established for $\mathcal{CH}(S, W)$ cannot be achieved by the concatenated codes, MSR convolutional codes are the only optimal codes, but huge finite fields are necessary to ensure their existence. To build a (n, k, μ) MSR convolutional code, where μ is the memory of the code, the field required for the construction presented in Mahmood et al. (2015) is $\mathbb{F}_{q^M}$ with $M = q^{n(\mu+2)-1}$. For example, to obtain the codes $(3, 1, 2)$ and $(2, 1, 2)$ the fields $\mathbb{F}_{2^{2048}}$ and $\mathbb{F}_{2^{128}}$ are needed.

This last condition is an issue from the practical point of view, since the memory required for storage and usage of these codes are enormous. For this reason, it makes sense to restrict the size of the fields for the construction of codes used in networks described as above. Taking this into account, we have two possibilities: a) to look for an alternative family of codes which achieves or improve the bounds of the Rank-Deficient Sliding Window

Network CH(S,W) over fields of minimum size or b) to find a construction for MSR codes for smaller fields.

With this in mind, an alternative family of codes for this sort of channel are the concatenated codes. Despite the fact that their distance is significant smaller that the one of MSR codes (as indicated in Theorem 1) it is important to note that the distribution of the deficiencies within the window is crucial for recovering the missing packets. For instance, suppose that in the first t instances we send $x_0, \ldots, x_t$ and received $y_0, \ldots, y_t$. In order to recover x_0 at time instant t with the MSR code, we need to have less than $d_{SR}^{j-1}(\mathcal{C}_{MSR})$ deficiencies for some $j = 0, 1, \ldots, t$ independently of the distribution of these deficiencies within the intervals. However, if we use a concatenated code then depending on the distribution of the deficiencies within the windows, the inner rank metric code gives very different erasures patterns to the outer code.

4. COMPLEXITY

In this section we compare the complexity of the decoding performance of both, the concatenated codes and the MSR convolutional codes. In Mahmood et al. (2015), it is said that the complexity of the decoding method, over a window of length $j-1$, for the MSR convolutional codes is $\mathcal{O}(((j-1)k)^3)$ over $\mathbb{F}_{q^M}$. This is due to the decoding method can be reduced to the inversion of a square matrix of this size.

In order to obtain the complexity of the decoding performance of the concatenated codes over a window of length $j-1$ we have to observe that it can be divide into two parts. First, the MRD code corrects the rank deficiency errors or give an erasure. Second, once all the packets are processed by the MRD code, the convolutional code corrects all the erasures in the window by solving a linear system. The performance of the MRD code over the window has complexity $\mathcal{O}((j-1)(k_I)^3)$ over $\mathbb{F}_{q^m}$, while the second part has $\mathcal{O}((n_I d_H^j(C_O))^2)$, since the convolutional code correct up to $d_H^j(C_O)$ packets that contains n_I elements over the field $\mathbb{F}_{q^{mk_I}}$. Thus the complexity of the decoding performance of the concatenated code is $\mathcal{O}((j-1)(k_I)^3)$ over $\mathbb{F}_{q^m}$. By considering, $M = m$ in order to compare comparable complexities, we have $\mathcal{O}(j(k_I)^3) \leq \mathcal{O}((jk)^3)$ which means that, exponentially, the decoding for the concatenated codes is faster.

5. CONCLUSION

In this paper we have discussed the difference in the performance of both MSR convolutional codes and concatenated codes under the same channel. We can say that the concatenated codes are faster by correcting the rank deficiencies and they correct a considerable amount of rank deficiencies with high probability. These codes can be constructed for fields of a more affordable size than the required for the existence of MSR codes.

There are some open problems that arise from this discussion. One of them is to find a construction of rank-metric convolutional codes with greater correction capability than the concatenated codes over a same size field or develop a construction for MSR codes for limited field size. In this sense, in Alfarano et al. (2020), a construction for optimal codes for Hamming column distance is presented. One last open question emerges, the search for new concatenated codes that improve the bounds presented in Napp et al. (2018) due to the flexibility of concatenation technique.

REFERENCES

Alfarano, G.N., Napp, D., Neri, A., and Requena, V. (2020). Weighted reed-solomon convolutional codes. *CoRR*, abs/2012.11417. URL `https://arxiv.org/abs/2012.11417`.

Almeida, P., Martínez-Peñas, U., and Napp, D. (2020). Systematic maximum sum rank codes. *Finite Fields and Their Applications*, 65, 101677.

Gabidulin, E. (1985). Theory of codes with maximum rank distance. *prob. Inf. Transm.*, 21, 1–12.

Gluesing-Luerssen, H., Rosenthal, J., and Smarandache, R. (2006). Strongly-mds convolutional codes. *IEEE Transactions on Information Theory*, 52, 584–598.

Ho, T., Médard, M., Kötter, R., Karger, C., Effros, M., Shi, J., and Leong, B. (2006). A random linear network coding approach to multicast. *IEEE Transactions on Information Theory*, 52, 413–430.

Johannesson, R. and Zigangirov, K.S. (1999). *Fundamentals of Convolutional Coding*. IEEE Press, New York.

Kötter, R. and Kschischang, F. (2008). Coding for error and erasures in random network coding. *IEEE Transactions on Information Theory*, 54, 3579–3591.

Mahmood, R., Badr, A., and Khisti., A. (2015). Convolutional codes with maximum column sum rank for network streaming. *2015 IEEE International Symposium on Information Theory (ISIT)*, 2271–2275.

McEliece, R.J. (1998). The algebraic theory of convolutional codes. In V. Pless and W. Huffman (eds.), *Handbook of coding theory*, volume 1, 1065–1138. Elsevier Science Publishers, The Netherlands.

Napp, D., Pinto, R., Rosenthal, J., and Santana, F. (2017a). Column rank distances of rank metric convolutional codes. In V.S. A. Barbero and O. Ytrehus (eds.), *Coding Theory and Applications*, 248–256. Springer International Publishing.

Napp, D., Pinto, R., Rosenthal, J., and Vettori, P. (2017b). MRD rank metric convolutional codes. *2017 IEEE International Symposium on Information Theory (ISIT)*, 2766–2770.

Napp, D., Pinto, R., and Sidorenko, V. (2018). Concatenation of convolutional codes and rank metric codes for multi-shot network coding. *Des. Codes Cryptogr*, 86, 303–318.

Wachter-Zeh, A., Stinner, M., and Sidorenko, V. (2015). Convolutional codes in rank metric with application to random network coding. *IEEE Transactions on Information Theory*, 61, 3199–3213.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Reset control analysis and design for hybrid Lur'e dynamical systems

Agustina D'Jorge * **Isabelle Queinnec** * **Sophie Tarbouriech** *
Luca Zaccarian *,**

* *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*
(e-mail: {adjorge|queinnec|tarbour|zaccarian@laas.fr)
** *Department of Industrial Engineering, University of Trento, Italy*

Abstract: We introduce a new class of hybrid Lur'e dynamical systems where a sector nonlinearity may affect both the continuous-time evolution and the reset map acting on suitable closed-loop states, under a time-regularization mechanism ensuring dwell time of solutions. For this class of systems we characterize Lyapunov-based exponential stability conditions exploiting homogeneity of the closed loop. In particular, we show that, with quadratic Lyapunov certificates these conditions can be cast as linear matrix inequalities. We then focus on the control design problem, where both the feedback gains acting on the continuous evolution and the reset action must be designed, in addition to the sets where such resets are triggered, expressed by sign-indefinite quadratic forms. For this control design problem we also show that the synthesis can be performed by solving a set of linear matrix inequalities.

Keywords: Reset control systems, Lur'e systems, Lyapunov theory, exponential stability

1. INTRODUCTION

Reset control systems (see, for example, Goebel et al. (2009, 2012), Prieur et al. (2018), Le and Teel (2021) and references therein) is a specific class of hybrid dynamical systems wherein continuous motion of the plant-controller dynamics is equipped with resetting rules inducing instantaneous re initializations of certain controller states, whenever the so-called reset conditions are satisfied. Beyond the fact that this class of systems allows to deal with a broad range of applications, as automotive systems, power systems and biological systems, they can overcome the limitation of classical continuous control law and achieve desired behavior as for example robustness, performance improvement (see, e.g., Hespanha and Morse (1999); Prieur (2005), Safaei et al. (2010), Aangenent et al. (2010); Beker et al. (2001); Goebel and Teel (2009); Hespanha et al. (2003); Prieur and Astolfi (2003); Nešić et al. (2008, 2011), Prieur et al. (2011, 2013)).

In this paper, a new class of hybrid Lur'e dynamical systems is introduced, where a sector nonlinearity may affect both the continuous-time evolution and the reset map acting on suitable closed-loop states. In other words, we consider reset control systems with Lur'e non-linearity, consisting in the interconnection between a linear plant and a non-linearity through a feedback loop. The nonlinearity verifies a cone bounded sector condition Khalil (2002), Castelan et al. (2008). Lyapunov-based exponential stability conditions exploiting homogeneity of the closed loop are proposed by adapting results issued from Goebel et al. (2012); Nešić et al. (2008); Zaccarian et al. (2011). Following the ideas presented in Fichera et al. (2016a), Fichera et al. (2016b), with quadratic Lyapunov certificates these conditions can be cast as linear matrix inequalities (LMIs). We then focus on the control design problem, where we must design 1) the state feedback gains acting on the continuous evolution and the reset action and 2) the shape of the sets where such resets are triggered, expressed as sign-indefinite quadratic forms. For this control design problem we also show that the synthesis can be performed by solving a set of linear matrix inequalities. The contribution of the current note can be viewed as complementary to the results developed in Fiacchini et al. (2012) dealing with quadratic stability problem for hybrid systems with nested saturations.

The extended abstract is organized as follows: Section 2 introduces the class of hybrid systems under consideration and states the problems at stake. Section 3 presents theoretical results dealing with the stability analysis. Section 4 then expands the stability analysis results in order to handle the control design problem. Finally, some concluding remarks are given in Section 5.

Notation. The notation is standard. The Euclidean norm of a vector is denoted by $|\cdot|$. If $\mathcal{A}$ is a compact set, the notation $|x|_{\mathcal{A}} = \min\{|x-y| : y \in \mathcal{A}\}$ indicates the distance of the vector x from the set $\mathcal{A}$. If $\mathcal{A}$ is the origin then $|x|_{\mathcal{A}} = |x|$. For any $s \in \mathbb{R}$, the function $\mathrm{dz} : \mathbb{R} \to \mathbb{R}$ is defined by $\mathrm{dz}(s) = 0$ if $|s| \leq 1$ and $\mathrm{dz}(s) = \mathrm{sign}(s)(|s|-1)$ if $|s| \geq 1$. Given a matrix Q, $\mathrm{He}(Q) = Q + Q^\top$. Moreover, $\lambda_{min}(Q)$ (respectively, $\lambda_{max}(Q)$) denotes the minimum (respectively, the maximum) eigenvalue of Q.

2. PROBLEM STATEMENT

Consider the following hybrid system, including an input nonlinearity ϕ,

$$\begin{cases} \dot{x} = A_F x + B_F \phi(u_F), \\ \dot{\tau} = 1 - \mathrm{dz}\left(\dfrac{\tau}{\rho}\right) \end{cases} \qquad (x,\tau) \in \mathcal{C} \qquad (1a)$$

$$\begin{cases} x^+ = A_J x + B_J \phi(u_J) \\ \tau^+ = 0, \end{cases} \qquad (x,\tau) \in \mathcal{D} \qquad (1b)$$

where $x \in \mathbb{R}^n$ is the physical state, $\tau \in \mathbb{R}$ is a dwell-time logic (with $\rho > 0$), and $u_F \in \mathbb{R}^m$ and $u_J \in \mathbb{R}^m$ are suitable

control inputs to be designed. The flow and jump sets $\mathcal{C}$ and $\mathcal{D}$ are defined as follows:

$$\begin{aligned}\mathcal{C} &:= \{(x,\tau) : x \in \mathcal{F} \text{ or } \tau \in [0,\rho]\} \\ &= \{(x,\tau) : x \in \mathcal{F}\} \cup \{(x,\tau) : \tau \in [0,\rho]\} \qquad (2)\\ \mathcal{D} &:= \{(x,\tau) : x \in \mathcal{J} \text{ and } \tau \in [\rho,2\rho]\} \\ &= \{(x,\tau) : x \in \mathcal{J}\} \cap \{(x,\tau) : \tau \in [\rho,2\rho]\} \qquad (3)\end{aligned}$$

with $\mathcal{F}$ and $\mathcal{J}$ symmetric cones defined by a symmetric (typically not sign definite) matrix $M = M^\top \in \mathbb{R}^{n\times n}$ as

$$\begin{aligned}\mathcal{F} &:= \{x \in \mathbb{R}^n : x^\top M x \le 0\} \\ \mathcal{J} &:= \{x \in \mathbb{R}^n : x^\top M x \ge 0\}.\end{aligned} \qquad (4)$$

The nonlinearity ϕ affecting the system input is a known, continuous, decentralized cone bounded nonlinearity (see, for example, Khalil (2002), Castelan et al. (2008)) as stated in the following assumption.

Assumption 1. The nonlinearity $\phi : \mathbb{R}^m \to \mathbb{R}^m$ is a known, continuous and decentralized function, which verifies the following generic cone bounded sector condition for any diagonal positive definite matrix $S \in \mathbb{R}^{m\times m}$:

$$\phi^\top(\zeta)\, S\, (\phi(\zeta) - \Omega\zeta) \le 0 \quad \forall \zeta \in \mathbb{R}^m \qquad (5)$$

Matrix $\Omega \in \mathbb{R}^{m\times m}$ is a positive definite diagonal matrix defining the sector $[0, \Omega_{i,i}]$, where each component ϕ_i of ϕ lies.

In this note, the inputs of the flow and jump maps are selected as linear static state feedbacks:

$$u_F = K_F x \text{ and } u_J = K_J x, \qquad (6)$$

with $K_F \in \mathbb{R}^{m\times n}$, $K_J \in \mathbb{R}^{m\times n}$. Then the closed loop (1)-(6) can be rewritten as

$$\begin{cases}\dot{x} = A_F x + B_F \phi(K_F x) \\ \dot{\tau} = 1 - \mathrm{dz}\left(\dfrac{\tau}{\rho}\right),\end{cases} \quad (x,\tau) \in \mathcal{C} \qquad (7a)$$

$$\begin{cases}x^+ = A_J x + B_J \phi(K_J x) \\ \tau^+ = 0,\end{cases} \quad (x,\tau) \in \mathcal{D}, \qquad (7b)$$

where the flow and jumps sets $\mathcal{C}$ and $\mathcal{D}$ are defined in equation (2) and (3), respectively.

Remark 1. Note that in the closed-loop system (7), the two nonlinearities $\phi(K_F x)$ and $\phi(K_J x)$ satisfy Assumption 1. Therefore, the generic relation (5) can be particularized for both $\phi(K_F x)$ and $\phi(K_J x)$ as follows

$$\phi(K_F x)^\top\, S_F\, (\phi(K_F x) - \Omega_F K_F x) \le 0 \quad \forall x \in \mathbb{R}^n \qquad (8)$$

$$\phi(K_J x)^\top\, S_J\, (\phi(K_J x) - \Omega_J K_J x) \le 0 \quad \forall x \in \mathbb{R}^n, \qquad (9)$$

which holds for any diagonal positive matrices $S_F \in \mathbb{R}^{m\times m}$, $S_J \in \mathbb{R}^{m\times m}$ and where $\Omega_F \in \mathbb{R}^{m\times m}$ and $\Omega_J \in \mathbb{R}^{m\times m}$ are suitable positive definite diagonal matrices. Matrices Ω_F and Ω_J are supposed to be known.

In this note we address 1) the stability analysis of system (1)-(6), or equivalently system (7), when K_F are K_J are given, and 2) the design problem where K_F are K_J must be designed.

These two complementary problems can be summarized as follows.

Problem 1. Given the gains K_F are K_J, devise conditions to guarantee that the compact set $\mathcal{A}$ defined as

$$\mathcal{A} = \{0\} \times [0, 2\rho] \subset \mathbb{R}^n \times [0, 2\rho] \qquad (10)$$

is globally asymptotically stable for the closed loop (1)-(6) (equivalently, system (7)).

Problem 2. Design the gains K_F are K_J in (4) such that the compact set $\mathcal{A}$ defined as in (10) is globally asymptotically stable for the closed loop (1)-(6) (equivalently system (7)).

3. STABILITY ANALYSIS RESULTS

In this section, theoretical results addressing Problem 1 are proposed by exploiting some ingredients provided in Fichera et al. (2016a), Prieur et al. (2018).

Recall that due to the dwell-time, the solutions (x,τ) to system (7) may flow outside the flow set as emphasized for example in Zaccarian et al. (2005),Prieur et al. (2018). Hence, to deal with the effects of dwell-time on trajectories and in order to allow for more design flexibility consider the following definitions,

$$\tilde{\mathcal{F}} = \{x \in \mathbb{R}^n : x^\top \tilde{M} x \le 0\} \qquad (11)$$

$$\tilde{\mathcal{F}}_\epsilon = \{x \in \mathbb{R}^n : x^\top \tilde{M} x - \epsilon x^\top x \le 0\} \qquad (12)$$

with $\tilde{M} = \tilde{M}^\top \in \mathbb{R}^{n\times n}$ and $\epsilon > 0$ to be designed.

Note that (12) is the ϵ-inflated version of (11), therefore the inclusion $\tilde{\mathcal{F}} \subset \tilde{\mathcal{F}}_\epsilon$ is always satisfied.

Stability conditions to solve Problem 1 are first proposed by focusing on a generic Lyapunov function, and then they are specialized to a quadratic Lyapunov function, thus leading to a convenient formulation involving convex linear matrix inequalities (LMI).

Theorem 1. Consider system (7) and the sets defined in (11) and (12). Assume that there exist a continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}_{\ge 0}$, positive real scalars $\alpha_1, \alpha_2, \alpha_3$ and a nonnegative scalar $\underline{\rho}$ satisfying

$$\alpha_1 |x|^2 \le V(x) \le \alpha_2 |x|^2, \qquad \forall x \in \mathbb{R}^n, \qquad (13)$$

$$\begin{aligned}&\langle \nabla V(x), A_F x + B_F \phi_F \rangle + \alpha_3 V(x) \\ &\quad - 2\phi_F^\top S_F (\phi_F - \Omega_F K_F x) < 0, \quad \forall x \in \tilde{\mathcal{F}}_\epsilon \setminus \{0\}\end{aligned} \qquad (14)$$

$$\begin{aligned}&V(A_J x + B_J \phi_J) - \exp(\alpha_3 \underline{\rho}) V(x) \\ &\quad - 2\phi_J^\top S_J (\phi_J - \Omega_J K_J x) \le 0, \quad \forall x \in \mathcal{J}\end{aligned} \qquad (15)$$

$$x^+ \in \tilde{\mathcal{F}}, \qquad \forall x \in \mathcal{J} \qquad (16)$$

$$\mathcal{F} \subset \tilde{\mathcal{F}}_\epsilon \qquad (17)$$

with ϕ_F and ϕ_J being shorthands for $\phi(K_F x)$ and $\phi(K_J x)$, respectively. Then there exists $\bar{\rho} > \underline{\rho}$ such that, for any $\rho \in (\underline{\rho}, \bar{\rho})$, the set $\mathcal{A} := \{0\} \times [0, 2\rho] \subset \mathbb{R}^n \times [0, 2\rho]$ is globally asymptotically stable for the hybrid closed-loop system (7).

Proof. The proof is omitted in this extended abstract. However, the key ingredients of the proof emerge from a careful expansion of the proof of Theorem 5.1 in Prieur et al. (2018) to the case of system (7) subject to the nonlinearities $\phi(K_F x)$ and $\phi(K_J x)$ satisfying the cone-bounded conditions (8) and (9), respectively. □

An interesting way to particularize Theorem 1 is to select a quadratic Lyapunov function $V(x) = x^\top P x$, in order to derive LMI-based conditions. Then, the following result can be stated in the context of Problem 1.

Proposition 1. Given System (7) and positive scalars $\alpha_3 > 0$, $\underline{\rho} > 0$, assume that there exist matrices $P = P^\top > 0$, $\tilde{M} = \tilde{M}^\top$, S_F, S_J and $\tilde{S}_J$ diagonal positive definite matrices, non-negative scalars τ_S, τ_R, $\tilde{\tau}_C$, $\tilde{\tau}_F$ and a positive scalar ϵ such that the following inequalities hold,

$$\begin{pmatrix}\mathrm{He}(PA_F) + \alpha_3 P - \tau_S(\tilde{M} - \epsilon I) & PB_F + K_F^\top \Omega_F S_F \\ B_F^\top P + S_F \Omega_F K_F & -2S_F\end{pmatrix} < 0 \qquad (18)$$

$$\begin{pmatrix} A_J^\top P A_J - \exp(\alpha_3 \rho) P + \tau_R M & A_J^\top P B_J + K_J^\top \Omega_J S_J \\ B_J^\top P A_J + S_J \Omega_J K_J & B_J^\top P B_J - 2 S_J \end{pmatrix} \leq 0 \tag{19}$$

$$\begin{pmatrix} A_J^\top \tilde{M} A_J + \tilde{\tau}_C M & A_J^\top \tilde{M} B_J + K_J^\top \Omega_J \tilde{S}_J \\ B_J^\top \tilde{M} A_J + \tilde{S}_J \Omega_J K_J & B_J^\top \tilde{M} B_J - 2 \tilde{S}_J \end{pmatrix} \leq 0 \tag{20}$$

$$\tilde{M} - \tilde{\tau}_F M \leq \epsilon I. \tag{21}$$

Then there exists $\bar{\rho} > \rho$ such that for any $\rho \in (\rho, \bar{\rho})$ the set $\mathcal{A}$ is globally asymptotically stable for the hybrid closed-loop system (7).

Proof. We only present a sketch of the proof, which consists in rewriting the conditions of Theorem 1 fpr the case where $V(x) = x^\top P x$. Indeed, let us consider the case of relation (14), which reads

$$\begin{aligned} 2x^\top P(A_F x + B_F \phi_F) + \alpha_3 x^\top P x \\ -2\phi_F^\top S_F(\phi_F - \Omega_F K_F x) < 0, \forall x \in \tilde{\mathcal{F}}_\epsilon \setminus \{0\} \end{aligned}$$

with ϕ_F the shorthands for $\phi(K_F x)$. We can handle the fact that the condition has to be satisfied for any $x \in \tilde{\mathcal{F}}_\epsilon \setminus \{0\}$ by using the S-procedure (see Boyd et al. (1994)) and the definition (12), which leads to

$$\begin{aligned} 2x^\top P(A_F x + B_F \phi_F) + \alpha_3 x^\top P x \\ -2\phi_F^\top S_F(\phi_F - \Omega_F K_F x) - -\tau_S x^\top (\tilde{M} - \epsilon I) x < 0 \end{aligned}$$

with $\tau_S \geq 0$. Hence, if relation (18) is verified then the above inequality and therefore condition (14) are satisfied.

The same reasoning holds for relation (15). In the case of condition (16), since ϕ_J is involved, one has to consider that this relation holds for ϕ_J satisfying the cone-bounded condition (9). This leads to relation (20). □

Remark 2. The inequalities presented in Proposition 1 are not LMIs in the matrix decision variables P, $\tilde{M}$, S_F, S_J, $\tilde{S}_J$ and the scalar decision variables τ_S, τ_R, $\tilde{\tau}_C$, $\tilde{\tau}_F$, ϵ due to the product between certain scalar and matrix decision variables. These products can be eliminated, thereby transforming the conditions into authentic LMIs, when multiplying by τ_S inequalities (20) and (21), and performing the following change of variables:

$$\bar{M} = \tau_S \tilde{M}, \bar{\epsilon} = \tau_S \epsilon, \bar{\tau}_C = \tau_S \tilde{\tau}_C, \bar{\tau}_F = \tau_S \tilde{\tau}_F, \tau_S \tilde{S}_J = \bar{S}_J.$$

so that the emerging conditions become linear in the transformed decision variables P, $\bar{M}$, S_F, S_J, $\bar{S}_J$, τ_R, $\bar{\tau}_C$, $\bar{\tau}_F$ and $\bar{\epsilon}$.

It still is required to fix a priori the two scalars α_3 and ρ but we emphasize that these quantities should be selected small enough for the construction to be effective and this helps in an iterative selection.

4. DESIGN RESULTS

In this section we take inspiration from Proposition 1 and Remark 2, to propose sufficient conditions to solve Problem 2. Recall that the matrices Ω_F and Ω_J involved in relations (8) and (9) are supposed to be known.

Proposition 2. Given System (7) and positive scalars $\alpha_3 > 0$ and $\rho > 0$, if there exist matrices $P = P^\top > 0$, $\bar{M} = \bar{M}^\top$, $\bar{K}_F$, $\bar{K}_J$, diagonal matrices $S_F > 0$ and $S_J > 0$, non-negative scalars $\bar{\tau}_F, \bar{\tau}_C, \tau_R \in \mathbb{R}_{\geq 0}$ and positive scalar $\bar{\epsilon}$ such that the following inequalities hold,

$$\begin{pmatrix} A_F^\top P + P A_F + \alpha_3 P - \bar{M} + \bar{\epsilon} I & P B_F + \bar{K}_F^\top \\ B_F^\top P + \bar{K}_F & -2 S_F \end{pmatrix} < 0 \tag{22}$$

$$\begin{pmatrix} A_J^\top P A_J - \exp(\alpha_3 \rho) P + \tau_R M & A_J^\top P B_J + \bar{K}_J^\top \\ B_J^\top P A_J + \bar{K}_J & B_J^\top P B_J - 2 S_J \end{pmatrix} \leq 0 \tag{23}$$

$$\begin{pmatrix} A_J^\top \bar{M} A_J + \bar{\tau}_C M & A_J^\top \bar{M} B_J + \bar{K}_J^\top \\ B_J^\top \bar{M} A_J + \bar{K}_J & B_J^\top \bar{M} B_J - 2 S_J \end{pmatrix} \leq 0 \tag{24}$$

$$\bar{M} - \bar{\tau}_F M \leq \bar{\epsilon} I, \tag{25}$$

Then there exists $\bar{\rho} > 0$ such that for any $\rho \in (\rho, \bar{\rho})$ the set $\mathcal{A}$ is globally exponentially stable for the hybrid closed-loop system (7) with the gains $K_F = S_F^{-1} \Omega_F^{-1} \bar{K}_F$ and $K_J = S_J^{-1} \Omega_J^{-1} \bar{K}_J$.

Proof. The proof follows the same lines as in the proof of Proposition 1 by modifying the conditions thanks to the change of variables suggested in Remark 2. Furthermore, one can also use the change of variables $\bar{K}_F = S_F \Omega_F K_F$ and $\bar{K}_J = S_J \Omega_J K_J$, allowing to recover the gains K_F and K_J because matrices S_F, S_J, Ω_F and Ω_J are all diagonal positive definite and therefore nonsingular. □

5. CONCLUSION

This note introduced a new class of hybrid Lur'e dynamical systems where a sector nonlinearity may affect both the continuous-time evolution and the reset map acting on suitable closed-loop states, under a time-regularization mechanism ensuring dwell time of solutions. For this class of systems, both the stability analysis and the control design problems have been addressed by exploiting Lyapunov-based stability conditions and homogeneity of the closed loop. By selecting a quadratic Lyapunov certificate these conditions can be cast as linear matrix inequalities. In the control design problem, both the feedback gains acting on the continuous evolution and the reset action can be designed. For this control design problem the synthesis can also be performed by solving a set of linear matrix inequalities.

The studies proposed are preliminary but pave the way for future directions of research including addressing regional stability properties and taking into account nonlinearities affecting also the shape of the flow and jump sets. Another interesting problem could be to study how the conditions change if the flow and jump maps are subject to different input-nonlinearities, with independent sector bounds.

ACKNOWLEDGEMENTS

This work has been supported in part by the ANR Labex CIMI (grant ANR-11-LABX- 0040) within the French State Program "Investissement d'Avenir", by ANR via project HANDY, number ANR-18-CE40-0010 and by IRT Tremplin HYQPI.

REFERENCES

Aangenent, W., Witvoet, G., Heemels, W., Van De Molengraft, M., and Steinbuch, M. (2010). Performance analysis of reset control systems. *International Journal of Robust and Nonlinear Control*, 20(11), 1213–1233.

Beker, O., Hollot, C., and Chait, Y. (2001). Plant with integrator: an example of reset control overcoming limitations of linear feedback. *IEEE Transactions on Automatic Control*, 46(11), 1797–1799. doi:10.1109/9.964694.

Boyd, S., Ghaoui, L.E., Feron, E., and Balakrishnan, V. (1994). *Linear Matrix Inequalities in System and Control Theory.* Society for Industrial and Applied Mathematics.

Castelan, E.B., Tarbouriech, S., and Queinnec, I. (2008). Control design for a class of nonlinear continuous-time systems. *Automatica*, 44(8), 2034–2039.

Fiacchini, M., Tarbouriech, S., and Prieur, C. (2012). Quadratic stability for hybrid systems with nested saturations. *IEEE Transactions on Automatic control*, 57(7), 1832–1838.

Fichera, F., Prieur, C., Tarbouriech, S., and Zaccarian, L. (2016a). LMI-based reset H∞ analysis and design for linear continuous-time plants. Technical report, Hal preprint, hal-01236972, https://hal.laas.fr/hal-01236972v1.

Fichera, F., Prieur, C., Tarbouriech, S., and Zaccarian, L. (2016b). LMI-based reset H∞ design for linear continuous-time plants. *IEEE Transactions on Automatic Control*, 61(12), 4157–4163. doi:10.1109/TAC.2016.2552059.

Goebel, R., Sanfelice, R., and Teel, A. (2009). Hybrid dynamical systems. *IEEE Control Systems Magazine*, 29(2), 28–93. doi:10.1109/MCS.2008.931718.

Goebel, R., Sanfelice, R., and Teel, A. (2012). *Hybrid dynamical systems*. Princeton University Press.

Goebel, R. and Teel, A. (2009). Direct design of robustly asymptotically stabilizing hybrid feedback. *ESAIM: Control, Optimisation and Calculus of Variations*, 15(1), 205–213.

Hespanha, J., Liberzon, D., and Morse, A. (2003). Hysteresis-based switching algorithms for supervisory control of uncertain systems. *Automatica*, 39(2), 263–272.

Hespanha, J. and Morse, A. (1999). Stabilization of nonholonomic integrators via logic-based switching. *Automatica*, 35(3), 385–393.

Khalil, H.K. (2002). *Nonlinear Systems, 3rd edition*. Patience Hall.

Le, J. and Teel, A. (2021). Passive soft-reset controllers for nonlinear systems. In *60th IEEE Conference n Decision and Control (CDC)*.

Nešić, D., Teel, A., and Zaccarian, L. (2011). Stability and performance of siso control systems with first-order reset elements. *IEEE Transactions on Automatic Control*, 56(11), 2567–2582.

Nešić, D., Zaccarian, L., and Teel, A. (2008). Stability properties of reset systems. *Automatica*, 44(8), 2019–2026.

Prieur, C. (2005). Asymptotic controllability and robust asymptotic stabilizability. *SIAM Journal on Control and Optimization*, 43(5), 1888–1912.

Prieur, C. and Astolfi, A. (2003). Robust stabilization of chained systems via hybrid control. *IEEE Transactions on Automatic Control*, 48(10), 1768–1772.

Prieur, C., Queinnec, I., Tarbouriech, S., and Zaccarian, L. (2018). Analysis and synthesis of reset control systems. *Foundations and Trends in Systems and Control*, 6(2-3), 117–338.

Prieur, C., Tarbouriech, S., and Zaccarian, L. (2011). Improving the performance of linear systems by adding a hybrid loop. In *IFAC WC 2011 - 18th IFAC World Congress*, p. 6301–6306. Milan, Italy.

Prieur, C., Tarbouriech, S., and Zaccarian, L. (2013). Lyapunov-based hybrid loops for stability and performance of continuous-time control systems. *Automatica*, 49(2), 577–584.

Safaei, F., Hespanha, J., and Stewart, G. (2010). Quadratic optimization for controller initialization in multivariable switching systems. In *Proceedings of the 2010 American Control Conference*, 2511–2516. doi:10.1109/ACC.2010.5530588.

Zaccarian, L., Nesic, D., and Teel, A. (2005). First order reset elements and the clegg integrator revisited. In *Proceedings of the 2005, American Control Conference, 2005.*, 563–568 vol. 1. doi:10.1109/ACC.2005.1470016.

Zaccarian, L., Nešić, D., and Teel, A. (2011). Analytical and numerical Lyapunov functions for SISO linear control systems with first-order reset elements. *International Journal of Robust and Nonlinear Control*, 21(10), 1134–1158.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Reduced Control Systems on Symmetric Lie Algebras ⋆

Emanuel Malvetti * **Gunther Dirr** ** **Frederik vom Ende** *
Thomas Schulte-Herbrüggen *

* *Dept. Chem., Lichtenbergstraße 4, 85747 Garching, Germany & Munich Centre for Quantum Science and Technology (MCQST), (e-mail: {emanuel.malvetti, frederik.vom-ende, tosh}@tum.de).*
** *Institute of Mathematics, University of Würzburg, Emil-Fischer-Straße 40, 97074 Würzburg, Germany, (e-mail: dirr@mathematik.uni-wuerzburg.de)*

date: 15 July 2022

Abstract: For a symmetric Lie algebra $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ we consider a class of bilinear or more general control-affine systems on $\mathfrak{p}$ defined by a drift vector field X and control vector fields ad_{k_i} which gain fast and full control on the adjoint orbits of the corresponding compact group $\mathbf{K}$. We show that under quite general assumptions on X such a control system is essentially equivalent to a natural reduced system on a maximal Abelian subspace $\mathfrak{a} \subseteq \mathfrak{p}$, and likewise to related differential inclusions defined on $\mathfrak{a}$.

Keywords: Control-affine Systems; Differential Inclusions; Symmetric Lie algebras; Adjoint Orbits

1. INTRODUCTION

We consider control systems that admit fast controllability on certain degrees of freedom, represented by a Lie group action. The goal is to define an associated (reduced) control system on the remaining degrees of freedom, and to show that the two systems are essentially equivalent (in a sense which will be specified later). Of course, this idea is not new and has been applied, e.g., in a seminal work by [Khaneja et al. (2001)] to compute time-optimal controls in certain low-dimensional closed quantum systems. The success of their approach is based on the symmetric space structure of the resulting quotient manifold, i.e. of the manifold which results from factoring out the orbits of the fast controllable dynamics. This method was further exploited by Khaneja and Yuan to characterise the reachable sets of certain bilinear systems, cf. [Yuan et al. (2018)].

In contrast, our subsequent framework is motivated by open quantum dynamics, where for certain experimental settings with switchable noise (i.e. fast controllable coupling to an environment [Chen et al. (2014); Wong et al. (2019)], described in [Bergholm et al. (2016)]) one can assume fast and full control on the unitary orbits in the state space of all density matrices. The drift vector field then usually describes relaxation or, more generally, interaction of the system with some environment. In [Dirr et al. (2019)] a simplified model for such systems was studied, where the state space could be restricted to diagonal density matrices (due to some invariance condition on the drift vector field) and unitary orbits therefore collapsed to a discrete Weyl group action (of permutation matrices). However, without the said invariance condition one is led to factor out the full unitary dynamics. This approach has been pursued in [Rooney et al. (2018)] yet under an additional boundedness condition, which is superfluous as we will show in Thm. 2.

Our general setting will be as follows: Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be a semisimple, orthogonal, symmetric Lie algebra defined by its Cartan-like decomposition. This means that $\mathfrak{g}$ is a semisimple Lie algebra, where $\mathfrak{k}$ and $\mathfrak{p}$ are the $+1$ and -1 eigenspaces of some involutive Lie algebra automorphism θ of $\mathfrak{g}$. We will set $\mathbf{K} := \mathrm{Int}_{\mathfrak{g}}(\mathfrak{k})$, the subgroup, generated by $\mathfrak{k}$, of the group of inner automorpisms of $\mathfrak{g}$. Then $\mathbf{K}$ is a compact Lie group and we can endow $\mathfrak{p}$ with a $\mathbf{K}$-invariant inner product. By $\mathfrak{a}$ we denote a choice of maximal Abelian subspace of $\mathfrak{p}$, and by $\mathbf{W} = N(\mathfrak{a})/Z(\mathfrak{a})$ the Weyl group acting on $\mathfrak{a}$, where $N(\mathfrak{a})$ and $Z(\mathfrak{a})$ denote the normalizer and centralizer of $\mathfrak{a}$ in $\mathbf{K}$. The Weyl group is a finite reflection group, and it admits a (closed) Weyl chamber $\mathfrak{w} \subset \mathfrak{a}$. One can show that $\mathfrak{w} \cong \mathfrak{a}/\mathbf{W} \cong \mathfrak{p}/\mathbf{K}$ are canonically isometric. For further details we refer to, e.g., [Helgason (1978)].

Example 1. (Eigenvalue decomposition). Consider the symmetric Lie algebra $\mathfrak{sl}(n, \mathbb{C}) = \mathfrak{su}(n) \oplus \mathfrak{herm}_0(n, \mathbb{C})$, where $\mathfrak{herm}_0(n, \mathbb{C})$ denotes the traceless Hermitian matrices. The corresponding Lie algebra automorphism is $\theta(x) := -x^*$. Then a convenient choice of a maximal Abelian subspace of $\mathfrak{herm}_0(n, \mathbb{C})$ is the set $\mathfrak{diag}_0(\mathbb{R})$ of real traceless diagonal matrices. The Weyl group is isomorphic to the symmetric group S_n acting by permutation of the diagonal entries of the elements of $\mathfrak{diag}_0(\mathbb{R})$, and a natural choice of a Weyl chamber is given by the real traceless diagonal matrices with diagonal elements in weakly decreasing order.

⋆ The project was supported i.a. by Excellence Network of Bavaria under ExQM and is part of *Munich Quantum Valley* of the Bavarian State Government with funds from Hightech Agenda *Bayern Plus*.

Example 2. Consider the symmetric Lie algebra $\mathfrak{su}(n) = \mathfrak{so}(n,\mathbb{R}) \oplus i\mathfrak{sym}_0(n,\mathbb{R})$ corresponding to the compact Riemannian symmetric space $\mathrm{SU}(n)/\mathrm{SO}(n)$. Here $\mathfrak{sym}_0(n,\mathbb{R})$ denotes the traceless real symmetric matrices. This symmetric Lie algebra is isomorphic to the symmetric Lie algebra $\mathfrak{sl}(n,\mathbb{R}) = \mathfrak{so}(n,\mathbb{R}) \oplus \mathfrak{sym}_0(n,\mathbb{R})$, which now corresponds to the non-compact Riemannian symmetric space $\mathrm{SL}(n,\mathbb{R})/\mathrm{SO}(n)$. Such a duality holds for all symmetric Lie algebras. Note that the latter example corresponds to the orthogonal diagonalization of (traceless) real symmetric matrices.

In fact many common matrix diagonalizations, e.g. the singular value decomposition, and also some uncommon ones, can be rephrased in the setting of symmetric Lie algebras [Kleinsteuber (2006)].

After these preliminaries we can define the class of control-affine systems on $\mathfrak{p}$ that we want to study in the sequel. Given a vector field $X \in \mathfrak{X}(\mathfrak{p})$, and a set of control directions $k_1, \dots, k_m \in \mathfrak{k}$ satisfying $\langle k_1, \dots, k_m \rangle_{\mathsf{Lie}} = \mathfrak{k}$, we consider the control system

$$\dot{p}(t) = X(p(t)) + \sum_{i=1}^{m} u_i(t)\,\mathrm{ad}_{k_i}(p(t)),\ p(0) = p_0 \in \mathfrak{p} \quad \text{(A)}$$

where ad_x denotes the adjoint operator of x, that is, $\mathrm{ad}_x(y) := [x,y]$. The control functions $u_i : [0,T] \to \mathbb{R}$ are required to be locally integrable. A solution $p : [0,T] \to \mathfrak{p}$ is an absolutely continuous function satisfying (A) almost everywhere. Obviously (A) reduces to a bilinear system [Elliott (2009)] if X is linear.

Since the control directions k_i generate the entire Lie algebra $\mathfrak{k}$, and since we do not impose any bounds (neither of L^∞ nor of L^1 type) on the control functions, we can move within the $\mathbf{K}$-orbits of $\mathfrak{p}$ from any starting point to any desired target point arbitrarily quickly [Jurdjevic and Sussmann (1972)]. We say that we have fast and full control on the Lie group $\mathbf{K}$ and consequently on any $\mathbf{K}$-orbit of $\mathfrak{p}$. In particular we can move into the maximal Abelian subspace $\mathfrak{a}$ at any time. This motivates us to define a reduced control system on $\mathfrak{a}$. To this end we first introduce some useful concepts.

For every $K \in \mathbf{K}$, we define the *induced vector field*

$$X_K := \Pi_{\mathfrak{a}} \circ \mathrm{Ad}^\star_K(X) \in \mathfrak{X}(\mathfrak{a}),$$

where $\Pi_{\mathfrak{a}}$ is the orthogonal projection in $\mathfrak{p}$ on $\mathfrak{a}$. By Ad_K we denote the adjoint action of $\mathbf{K}$ on $\mathfrak{p}$ and by $\mathrm{Ad}^\star_K$ its pullback action, i.e.,

$$\mathrm{Ad}^\star_K(X) = \mathrm{Ad}_K^{-1} \circ X \circ \mathrm{Ad}_K\,.$$

If X is linear, then so are all X_K. Furthermore we define the set of *achievable derivatives* at $a \in \mathfrak{a}$ by

$$\Lambda(a) = \{X_K(a) : K \in \mathbf{K}\} \subset T_a\mathfrak{a} \cong \mathfrak{a}.$$

Since $\mathbf{K}$ is compact, $\Lambda(a)$ is also compact for all a. We can interpret $\Lambda : \mathfrak{a} \to \mathcal{P}(\mathfrak{a})$ as a set valued function. It turns out that if X is Lipschitz, e.g. if it is linear, then the set valued function Λ is also Lipschitz. This means that for all $x, y \in \mathfrak{a}$,

$$\Lambda(x) \subseteq \Lambda(y) + L\|x-y\|B_1$$

for some (global) Lipschitz constant $L > 0$ and where B_1 denotes the closed unit ball in $\mathfrak{a}$.

Now we can define the *reduced control* system by

$$\dot{a}(t) = X_{U(t)}(a(t)), \quad a(0) = a_0 \in \mathfrak{a}, \qquad \text{(Q)}$$

where the control function $U : [0,T] \to \mathbf{K}$ has to be measurable [1]. Again, a solution $a : [0,T] \to \mathfrak{a}$ is absolutely continuous and satisfies (Q) almost everywhere. We can also define a corresponding *differential inclusion*

$$\dot{a}(t) \in \Lambda(a(t)), \quad a(0) = a_0 \in \mathfrak{a}, \qquad \text{(I)}$$

where a solution $a : [0,T] \to \mathfrak{a}$ needs to be absolutely continuous and satisfy (I) almost everywhere. In fact, if X is continuous, (Q) and (I) are equivalent, i.e. they have the same solutions, see Thm. 2.3 in [Smirnov (2002)]. Often it will be convenient to consider a relaxed version of the differential inclusion given by

$$\dot{a}(t) \in \mathrm{conv}(\Lambda(a(t))), \quad a(0) = a_0 \in \mathfrak{a}, \qquad \text{(R)}$$

where conv denotes the convex hull. This will slightly enlarge the set of solutions, however, if X is Lipschitz, every solution of (R) can still be approximated uniformly (on compact time intervals) by solutions to (I), see Ch. 2.4, Thm. 2 in [Aubin et al. (1984)].

If X is Lipschitz, then Λ is Lipschitz with compact values, which implies some convenient properties of the relaxed control system (R), see Ch. 4 in [Smirnov (2002)].

Proposition 1. Let X be Lipschitz and let $a_0 \in \mathfrak{a}$. It holds that:

(i) The set $S_{[0,T]}(a_0)$ of solutions to (R) is path connected in the AC-topology [2];
(ii) If Λ is bounded, then $S_{[0,T]}(a_0)$ is compact in the standard C-topology of uniform convergence;
(iii) If $a \in S_{[0,T]}(a_0)$ is a solution to (R) with $a(T) \in \partial\mathsf{reach}_{[0,T]}(a_0)$, then $a(t) \in \partial\mathsf{reach}_{[0,t]}(a_0)$ for all $t \in [0,T]$;
(iv) If Λ is bounded, then there exist time-optimal solutions to (R) starting in a given compact set and ending in a given closed set, if any such solution exists;
(v) If X is Lipschitz with global Lipschitz constant L, then the map $S_{[0,T]} : \mathfrak{a} \to AC([0,T],\mathfrak{a})$ is Lipschitz with global Lipschitz constant $1 + TLe^{TL}$.

2. MAIN RESULTS

Our main results describe the equivalence of the control-affine system (A) on $\mathfrak{p}$ and the reduced control system (Q) on $\mathfrak{a}$. Detailed proofs will be presented in the MTNS-talk and published in a subsequent journal paper. Here – due to page constraints – we focus on a sketch of the key ideas.

First we show a local equivalence result which illustrates why the definition of the reduced control system is reasonable. For this we need the following fact: if $p : [0,T] \to \mathfrak{p}$ is differentiable at $t \in [0,T]$, then there is some $a : [0,T] \to \mathfrak{a}$ differentiable at t satisfying $\pi \circ p = \pi \circ a$. Furthermore this derivative is well defined up to a Weyl group action on the pair $(a(t), \dot{a}(t))$. [3] Hence we can define the set

[1] Due to the compactness of $\mathbf{K}$, the control function U is automatically bounded, and hence in L^∞.

[2] By $AC([0,T],\mathfrak{a})$ we denote the Banach space of absolutely continuous functions $a : [0,T] \to \mathfrak{a}$ with the norm $\|a\|_{AC} = |a(0)| + \int_0^T |\dot{a}(t)|dt$.

[3] In the case of unitary diagonalization of Hermitian matrices (Example 1), this is a well-known result, see Ch. I.§5, Thm. 1 in [Rellich (1969)]. We extended the result to all semisimple, orthogonal, symmetric Lie algebras.

$$\tilde{\Lambda}(a) = \{\dot{b}(t) : b : [0,T] \to \mathfrak{a} \text{ differentiable at } t, b(t) = a, \\ \pi \circ b = \pi \circ p, \text{ and } p : [0,T] \to \mathfrak{p} \text{ solves (A)}\},$$

which is the set of all possible derivatives in $\mathfrak{a}$ of solutions to (A). Recall that $\mathbf{K}_a$ and $\mathbf{W}_a$ denote the stabilizers of a in $\mathbf{K}$ and $\mathbf{W}$ respectively.

Proposition 2. (Local equivalence). It holds that

$$\tilde{\Lambda}(a) \subseteq \Lambda(a) = \bigcup_{v \in \tilde{\Lambda}(a)} \operatorname{conv}(\mathbf{W}_a v) \subseteq \operatorname{conv}(\tilde{\Lambda}(a)),$$

and in particular,

$$\operatorname{conv}(\Lambda(a)) = \operatorname{conv}(\tilde{\Lambda}(a)).$$

Proof. [Sketch of proof] The first inclusion follows from a computation as in the proof of the following theorem. For the first equality a computation shows that (we replace a by b for clarity)

$$\begin{aligned}\Lambda(b) &= \bigcup_{[K] \in \mathbf{K}/\mathbf{K}_b} \{(\Pi_{\mathfrak{a}} \circ \operatorname{Ad}_{\tilde{K}}^{-1} \circ \operatorname{Ad}_K^\star(X))(b) : \tilde{K} \in \mathbf{K}_b\} \\ &= \bigcup_{[K] \in \mathbf{K}/\mathbf{K}_b} \operatorname{conv}(\mathbf{W}_b \cdot (\pi_b \circ \Pi_b \circ \operatorname{Ad}_K^\star(X))(b)),\end{aligned}$$

where the second transformation uses Kostant's convexity theorem applied to the symmetric Lie subalgebra of $\mathfrak{g}$ given by the commutant of b. Here Π_b denotes the orthogonal projection onto the commutant of b in $\mathfrak{p}$, and $\pi_b : \mathfrak{p}_b \to \mathfrak{p}_b/\mathbf{K}_b$ denotes the quotient map. The rest is straightforward.

Now we will prove a global equivalence result. The first direction, projecting from $\mathfrak{p}$ to $\mathfrak{a}$ is the easier one. In the following we denote by $\pi : \mathfrak{p} \to \mathfrak{w} \cong \mathfrak{p}/\mathbf{K}$ the quotient map, where $\mathfrak{w}$ is some choice of Weyl chamber.

Theorem 1. Let $p : [0,T] \to \mathfrak{p}$ be a solution to the control-affine system (A). Then $a = \pi \circ p$ is a solution to the reduced control system (Q).

Proof. [Sketch of proof] The quotient map π is non-expansive, and hence a is absolutely continuous. On the subset $J \subseteq [0,T]$ where both p and a are differentiable, it holds that there is some $K : J \to \mathbf{K}$ such that $a = \operatorname{Ad}_K^{-1}(p)$ and $\dot{a} = \operatorname{Ad}_K^{-1} \circ \Pi_p(\dot{p})$. Then a computation shows that $\dot{a} = X_K(a)$ on J, and hence (as J clearly has full measure) a satisfies the differential inclusion (I) almost everywhere.

More delicate is the other direction, lifting from $\mathfrak{a}$ to $\mathfrak{p}$.

Theorem 2. Assume that X is Lipschitz and real analytic. Let $a : [0,T] \to \mathfrak{a}$ be a solution to the reduced control system (Q) with control function $U : [0,T] \to \mathbf{K}$. Then, for every $\varepsilon > 0$ there exists a solution $p : [0,T] \to \mathfrak{p}$ to the control-affine system (A) satisfying $\|\operatorname{Ad}_k(a) - p\|_\infty \leq \varepsilon$. In particular $\|\pi \circ p - \pi \circ a\|_\infty \leq \varepsilon$.

Proof. [Sketch of proof] Using standard L^1-approximation results we may assume that U is real analytic and that $a(0)$ is regular [4], see Sec. 2.8 Thm. 1 in [Sontag (1998)]. Hence a is real analytic and regular with finitely many exceptions. Now we define an ideal lift q of a by $q = \operatorname{Ad}_U(a)$. Moreover we define an approximating solution p by $\dot{p} = (\operatorname{ad}_h + X)p$ and $p(0) = q(0)$, where h contains the derivative of U as well as the part of $X(q)$ tangential to the $\mathbf{K}$ orbit, outside of an ε region around the singular points of a. By definition h is bounded and hence p is well-defined. Then one can use Gronwall's inequality to bound the difference $p - q$ and one finds that p approximates q uniformly as $\varepsilon \to 0$. Finally, by standard techniques from control theory, cf. [Liu (1997)], one can uniformly approximate p using solutions to (A).

The problems occur at the points where the stabilizers in $\mathbf{K}$ or $\mathbf{W}$ respectively are not minimal. When projecting, this generally reduces the regularity. To deal with this we generalized several known results for Hermitian matrices to the setting of semisimple, orthogonal, symmetric Lie algebras. In particular, although the quotient map $\pi : \mathfrak{p} \to \mathfrak{w}$ is not differentiable at the singular points, it still preserves absolute continuity, and at the points where the projected path is differentiable, we can give an explicit expression for the derivative. When lifting, the control functions might need to diverge at these points. Intuitively this happens because the orbits of the corresponding stabilizers become arbitrarily small, and so the control in these directions becomes increasingly ineffective. In [Rooney et al. (2018)] the control system on the eigenvalue simplex was defined with the additional condition that such a divergence was not allowed to occur. We showed that this assumption is unnecessary, by approximating the solutions in a neighborhood of the singular points.

Example 3. To see that the approximation cannot be avoided in general, consider a system where $X(0) \neq 0$. Then $p \equiv 0$ is not a solution of (A), but $a \equiv 0$ is a solution of (Q). Indeed, by Kostant's convexity theorem [Kostant (1973)], and assuming that $X(0) \in \mathfrak{a}$, we see that

$$\Lambda(0) = \{\Pi_{\mathfrak{a}} \circ \operatorname{Ad}_K(X(0)) : K \in \mathbf{K}\} = \operatorname{conv}(\mathbf{W}(X(0)))$$

and hence $\Lambda(0)$ contains the convex combination

$$\frac{1}{|\mathbf{W}|} \sum_{w \in \mathbf{W}} w \cdot X(0),$$

which equals 0, the unique fixed point of a Weyl group action. Thus $a \equiv 0$ is a solution to (Q).

Finally, the above findings suggest to formalize the notion of equivalence [5] of control systems as follows:

Definition 1. Let X and Y be two (state) spaces on which are defined two control systems whose solutions are absolutely continuous functions. Let $\pi : X \to Y$ be a (at least Lipschitz-continuous) surjection. We say that the two control systems are equivalent if

(i) whenever $x : [0,T] \to X$ is a solution on X, then $\pi \circ x$ is a solution on Y;
(ii) whenever $y : [0,T] \to Y$ is a solution on Y, there exist solutions $x_n : [0,T] \to X$ such that $\pi \circ x_n$ uniformly approximates y as $n \to \infty$.

Then the equivalence between the control-affine and the reduced system follows immediately from Theorem 1 and Theorem 2. Moreover, their proofs reveal the following equivalence between their reachable sets:

Theorem 3. Assume that X is Lipschitz and real analytic. Then for each $T > 0$ and $p \in \mathfrak{p}$ one has

[4] Recall that $a \in \mathfrak{a}$ is called regular if it has trivial stabilizer in W, or equivalently, if it lies in the interior of a Weyl chamber.

[5] Note that this definition does not yield an equivalence relation.

$$\mathsf{reach}_T(p) \subseteq \mathrm{Ad}_{\mathbf{K}}(\mathsf{reach}_T(\pi(p))) \subseteq \overline{\mathsf{reach}_T(p)},$$

where the reachable sets refer the the control-affine system (A) on $\mathfrak{p}$ and the relaxed control system (R) on $\mathfrak{a}$.

3. CONCLUSION

We have shown that for a class of control systems defined on symmetric Lie algebras and with fast control on the corresponding compact Lie group (examples of which appear naturally in quantum control theory), one can define a reduced control system whose state space has a much lower dimension. The equivalence theorem shows that this reduction does not incur any loss of information. Hence this method yields a new perspective on these control systems, and in some cases allows to visualize properties of the system due to the reduction in dimension. Furthermore the new control system can be represented by a differential inclusion with nice properties, and thus allows for many new methods to be applied to the system.

REFERENCES

Aubin, J.P., Cellina, A. (1984). *Differential Inclusions: Set-Valued Maps and Viability Theory.* Springer, Berlin.

Bergholm, V., Wilhelm, F., and Schulte-Herbrüggen, T. (2016). Arbitrary n-Qubit State Transfer Implemented by Coherent Control and Simplest Switchable Local Noise. `https://arxiv.org/abs/1605.06473v2`.

Chen, Y., Neill, C., Roushan, P., Leung, N., Fang, M., Barends, R., Kelly, J., Campbell, B., Chen, Z., Chiaro, B., Dunsworth, A., Jeffrey, E., Megrant, A., Mutus, J.Y., O'Malley, P.J.J., Quintana, C.M., Sank, D., Vainsencher, A., Wenner, J., White, T.C., Geller, M.R., Cleland, A.N., and Martinis, J.M. (2014). Qubit Architecture with High Coherence and Fast Tunable Coupling. *Phys. Rev. Lett*, 113, 220502.

Dirr, G., vom Ende, F., and Schulte-Herbrüggen, T. (2019). Reachable Sets from Toy Models to Controlled Markovian Quantum Systems. *Proc. IEEE Conf. Decision Control (IEEE-CDC)*, 58, 2322. `https://arxiv.org/abs/1905.01224`.

Elliott, D. (2009). *Bilinear Control Systems: Matrices in Action.* Springer, London.

Helgason, S. (1978). *Differential Geometry, Lie Groups, and Symmetric Spaces.* Academic Press, New York.

Jurdjevic, V. and Sussmann, H. (1972). Control Systems on Lie Groups. *J. Diff. Equat.*, 12, 313–329.

Khaneja, N., Brockett, R., and Glaser, S.J. (2001). Time Optimal Control in Spin Systems. *Phys. Rev. A 63*, 032308.

Kleinsteuber, M. (2006). Jacobi-type Methods on Semisimple Lie Algebras – a Lie Algebraic Approach to the Symmetric Eigenvalue Problem. Ph.D. thesis, Department of Mathematics, University of Würzburg. `https://www.ei.tum.de/fileadmin/tueifei/ldv/gol/preprints/klei-phd06.pdf`.

Kostant, B. (1973). On Convexity, the Weyl Group and the Iwasawa Decomposition *Ann. Sci. Ecole Norm. Sup.*, 6, 413-455.

Liu, W. An Approximation Algorithm for Nonholonomic Systems. *SIAM J. Control Optim.* 35, 1328–1365 .

Rellich, F. (1969) *Pertubation Theory of Eigenvalue Problems.* Gordon and Breach, New York.

Rooney, P., Bloch, A., and Rangan, C. (2018). Steering the Eigenvalues of the Density Operator in Hamiltonian-Controlled Quantum Lindblad Systems. *IEEE Trans. Automat. Control*, 63, 672–681.

Smirnov, G. (2002). *Introduction to the Theory of Differential Inclusions.* Amer. Math. Soc., Providence, Rhode Island.

Sontag, E. (1998). *Mathematical Control Theory.* Springer, New York.

Wong, C., Wilen, C., McDermott, R., and Vavilov, M. (2019). A Tunable Quantum Dissipator for Active Resonator Reset in Circuit QED. *Quant. Sci. Technol.*, 4, 025001.

Yuan, H. and Khaneja, N. (2005). Reachable Set of Bilinear Control Systems With Time Varying Drift. *Proceedings of the 44th IEEE Conference on Decision and Control*, 8006-8011.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Asymptotic stability and structure preserving discretization for gas flow in networks

Herbert Egger * **Jan Giesselmann** ** **Teresa Kunkel** **
Nora Philippi **

* *Johannes Kepler University and Johann Radon Institute for Computational and Applied Mathematics, Linz, Austria (e-mail: herbert.egger@jku.at).*
** *TU Darmstadt, Germany (e-mail: giesselmann@mathematik.tu-darmstadt.de,tkunkel@mathematik.tu-darmstadt.de,philippi@mathematik.tu-darmstadt.de)*

Abstract: Gas transport in one-dimensional pipe networks can be described as an abstract dissipative Hamiltonian system, for which quantitative stability bounds can be derived by means of relative energy estimates for subsonic flow. This allows to establish convergence to the parabolic limit problem in the practically relevant high friction regime. The stability estimates carry over almost verbatim to a mixed finite element approximations with an implicit Euler time discretization, leading to order optimal convergence rates that are uniform the high friction limit. All results are proven in detail for the flow on a single pipe, but by the port-Hamiltonian formalism, they naturally extend to pipe networks.

Keywords: port-Hamiltonian systems, relative energy estimates, asymptotic preserving schemes

1. INTRODUCTION

We consider dynamical systems modelling the transport of gas in long pipes and pipe networks. On each pipe, identified with the interval $(0,\ell)$, the flow is modelled by the barotropic Euler equations which after transformation and rescaling are given by

$$a\partial_\tau \rho + \partial_x(a\rho w) = 0, \tag{1}$$

$$\varepsilon^2 \partial_\tau w + \partial_x\big(\frac{\varepsilon^2}{2} w^2 + P'(\rho)\big) = -\gamma |w| w. \tag{2}$$

Here a is the cross-sectional area of the pipe, ρ the gas density, τ, w and γ the rescaled time, velocity and friction coefficient, $P(\rho)$ the pressure potential, and ε a scaling parameter, proportional to the Mach number.

Of particular interest for the operation of gas transportation networks is the low-Mach, respectively, high friction limit $\varepsilon \to 0$, which describes to the practically relevant setting of long length and time scales; see Brouwer et al. (2011) for details. In this case, one can expect smooth solutions bounded away from vacuum, which is important in practice and therefore assumed in the following.

By formally setting $\varepsilon = 0$ in the system (1)–(2), and eliminating the velocity w via (2), one obtains a doubly nonlinear parabolic model for gas transport which is widely used in practice and which has also been studied intensively in the mathematical literature. Existence of weak solutions, uniform bounds, and converging discretization methods have been established for this models; see e.g. Bamberger et al. (1979); Schöbel-Kröhn (2020).

⋆ Supported by DFG via Grant TRR 154, projects C04 and C05.

2. ABSTRACT HAMILTONIAN FORMULATION.

The functions ρ, w in (1)–(2) are the *state variables* of the system, and we define corresponding *co-state variables* by

$$h := \tfrac{\varepsilon^2}{2} w^2 + P'(\rho), \qquad m := a\rho w, \tag{3}$$

which have the physical interpretation of the total specific enthalpy and mass flux, respectively. These auxiliary functions are linked via $ah = \frac{\delta \mathcal{H}}{\delta \rho}$ and $\varepsilon^2 m = \frac{\delta \mathcal{H}}{\delta w}$ to the variational derivatives of the associated energy functional

$$\mathcal{H}(\rho, w) := \int_0^\ell a\big(\frac{\varepsilon^2}{2}\rho w^2 + P(\rho)\big)\, dx. \tag{4}$$

Multiplying (1)–(2) with suitable test functions, integrating over $(0,\ell)$, and applying integration-by-parts in the second equation leads to

$$(a\partial_\tau \rho, q) + (\partial_x m, q) = 0, \tag{5}$$

$$(\varepsilon^2 \partial_\tau w, r) - (h, \partial_x r) = -(\gamma |w| w, r) - hr\big|_0^\ell, \tag{6}$$

which hold for all smooth test functions q and v, and for all points in time. We use $(u,v) = \int_0^\ell uv\, dx$ for abbreviation. By construction, any smooth solution of (1)–(2) satisfies these variational identities.

The system (5)–(6) can further be written as an abstract dissipative Hamiltonian system of the form

$$\mathcal{C}\partial_\tau \boldsymbol{u} + \big(\mathcal{J} + \mathcal{R}(\boldsymbol{u})\big)\boldsymbol{z}(\boldsymbol{u}) = \mathcal{B}_\partial \boldsymbol{z}(\boldsymbol{u}), \tag{7}$$

with $\boldsymbol{u} = (\rho, w)$ and $\boldsymbol{z}(\boldsymbol{u}) = \mathcal{C}^{-1}\mathcal{H}'(\boldsymbol{u}) = (h, m)$ denoting the state and co-state variables, and with appropriate

operators $\mathcal{C}$, $\mathcal{J}$, $\mathcal{R}$, $\mathcal{B}_\partial$. Note that $\mathcal{C}$ is positive definite, $\mathcal{J}$ skew-symmetric, and $\mathcal{R}(\boldsymbol{u})$ positive. From the abstract form of the problem, we immediately deduce the following energy identity or inequality for smooth solutions of (7)

$$\begin{aligned}\frac{d}{d\tau}\mathcal{H}(\boldsymbol{u}) &= \langle \partial_\tau \boldsymbol{u}, \mathcal{H}'(\boldsymbol{u})\rangle \\ &= -\langle \mathcal{R}(\boldsymbol{u})\boldsymbol{z}(\boldsymbol{u}), \boldsymbol{z}(\boldsymbol{u})\rangle + \langle \mathcal{B}_\partial \boldsymbol{z}(\boldsymbol{u}), \boldsymbol{z}(\boldsymbol{u})\rangle\end{aligned} \tag{8}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality product. This shows that the change in the total energy of the system over time is caused only by dissipation due to friction at pipe walls and by flux over the boundary.

3. ASYMPTOTIC STABILITY

In the following, we further investigate the dependence of solutions to (1)–(2) on the scaling parameter. To do so, we use the concept of *relative energy*, which is defined by

$$\mathcal{H}(\boldsymbol{u}|\hat{\boldsymbol{u}}) = \mathcal{H}(\boldsymbol{u}) - \mathcal{H}(\hat{\boldsymbol{u}}) - \langle \mathcal{H}'(\hat{\boldsymbol{u}}), \boldsymbol{u} - \hat{\boldsymbol{u}}\rangle. \tag{9}$$

For appropriately bounded subsonic states $\boldsymbol{u}$ and $\hat{\boldsymbol{u}}$, the relative energy introduces a distance measure which is equivalent to an ε-weighted L^2-norm; see Dafermos (2005); Egger and Giesselmann (2020) for details.

Now let $\boldsymbol{u} = (\rho, w)$, $\hat{\boldsymbol{u}} = (\hat\rho, \hat w)$ be solutions of (1)–(2) for different scaling parameters ε, $\hat\varepsilon$. Further assume that the pressure potential $P(\rho)$ is smooth and strictly convex, that the flow subsonic, and that the solutions are uniformly bounded and sufficiently smooth. Then we can show that

$$\begin{aligned}&\|\rho(\tau) - \hat\rho(\tau)\|^2_{L^2(0,\ell)} + \varepsilon^2 \|w(\tau) - \hat w(\tau)\|^2_{L^2(0,\ell)} \\ &\quad + \int_0^\tau \|w(s) - \hat w(s)\|^3_{L^3(0,\ell)} \le C e^{c\tau} |\varepsilon^2 - \hat\varepsilon^2|\end{aligned} \tag{10}$$

with constants c, C only depending on bounds for the solutions and parameters; see Egger and Giesselmann (2020) for the details. Let us note that this result holds in particular for $\hat\varepsilon = 0$, and thus yields asymptotic convergence towards the parabolic limit problem.

4. ASYMPTOTIC PRESERVING DISCRETIZATION

For discretization of the variational identities (5)–(6), we use a mixed finite element method in space together with an implicit Euler method in time. Then ρ_h^n and m_h^n denote the piecewise constant resp. piecewise linear approximations for ρ and m, while $w_h = w(\rho_h, m_h)$ and $h_h = h(\rho_h, m_h)$ are defined explicitly as functions of these functions. This amounts to a generalization of standard approximation schemes for related linear wave propagation problems Joly (2003). By formally setting $\varepsilon = 0$, we also obtain a viable numerical method for the parabolic limit problem, i.e., the scheme is *asymptotic preserving*.

Since the abstract Hamiltonian structure (7) of the problem is inherited by this variational discretization approach, also the stability analysis via relative energy estimates transfers directly. This allows us to prove error estimates

$$\begin{aligned}&\|\rho(\tau^n) - \rho_h^n\|^2_{L^2(0,\ell)} + \varepsilon^2 \|m(\tau^n) - m_h^n\|^2_{L^2(0,\ell)} \\ &\quad + \sum_{k=1}^n \Delta\tau \|m(\tau^k) - m_h^k\|^3_{L^3(0,\ell)} \le C(\Delta\tau^2 + h^2)\end{aligned} \tag{11}$$

with a uniform constant C being independent of ε; see Egger et al. (2021) for details. Let us emphasize that the error estimate remains valid in the asymptotic limit $\varepsilon = 0$ and thus also covers the parabolic limit problem. Further note that similar arguments were also employed for the analysis of numerical methods for the compresible Navier-Stokes equations; see Feireisl et al. (2018).

5. EXTENSION TO NETWORKS

Based on the variational framework employed in our analysis, all results presented for a single pipe can be generalized quite naturally to pipe networks, which are described by finite, directed and connected graphs. On every edge of the graph, representing a pipe of the network, the equations (1)–(2) are assumed to hold. Additional coupling conditions now have to be imposed in order to guarantee conservation of mass and energy across pipe junctions; see Reigstad (2015) for details. The weak formulation of this problem again leads to a system of the abstract form (7), such that all stability and convergence results derived for a single pipe carry over to the network almost immediately. Also the numerical scheme as well as its stability and convergence analysis via relative energy estimates hold almost verbatim.

ACKNOWLEDGEMENTS

This work was supported by the German Science Foundation (DFG) via grant TRR 154, projects C04 and C05.

REFERENCES

Bamberger, A., Sorine, M., and Yvon, J.P. (1979). Analyse et contrôle d'un réseau de transport de gaz. In *Computing methods in applied sciences and engineering*, volume 91 of *Lecture Notes in Phys.*, 347–359. Springer.

Brouwer, J., Gasser, I., and Herty, M. (2011). Gas pipeline models revisited: model hierarchies, nonisothermal models, and simulations of networks. *Multiscale Model. Simul.*, 9, 601–623.

Dafermos, C.M. (2005). *Hyperbolic conservation laws in continuum physics*. Springer.

Egger, H., Giesselmann, J., Kunkel, T., and Philippi, N. (2021). An asymptotic-preserving discretization scheme for gas transport in pipe networks. *arXiv:2108.13689*. Accepted for publication in IMA J. Numer. Anal.

Egger, H. and Giesselmann, J. (2020). Stability and asymptotic analysis for instationary gas transport via relative energy estimates. *arXiv:2012.14135*.

Feireisl, E., Lukacova-Medvidova, M., Necasova, S., Antonin, N., and She, B. (2018). Asymptotic preserving error estimates for numerical solutions of compressible Navier–Stokes equations in the low Mach number regime. *Multiscale Model. Simul.*, 16, 150–183.

Joly, P. (2003). Variational methods for time-dependent wave propagation problems. In *Topics in computational wave propagation*, volume 31 of *Lect. Notes Comput. Sci. Eng.*, 201–264. Springer, Berlin.

Reigstad, G.A. (2015). Existence and uniqueness of solutions to the generalized Riemann problem for isentropic flow. *SIAM J. Appl. Math.*, 75, 679–702.

Schöbel-Kröhn, L. (2020). *Analysis and numerical approximation of nonlinear evolution equations on network structures*. Dr. Hut Verlag, München.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On the Joint Spectral Radius of Shuffled Switched Linear Systems

Georges Aazan * **Antoine Girard** ** **Luca Greco** **
Paolo Mason **

* *Université Paris-Saclay, CNRS, ENS Paris-Saclay, Laboratoire Méthodes Formelles, 91190, Gif-sur-Yvette, France. (e-mail: georges.aazan@lsv.fr).*
** *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France. (e-mail: [firstname].[lastname]@l2s.centralesupelec.fr).*

Abstract: We present and develop tools to analyze stability properties of discrete-time switched linear systems driven by shuffled switching signals. A switching signal is said to be shuffled if all modes of the system are activated infinitely often. We establish a notion of joint spectral radius related to these systems: the shuffled joint spectral radius (SJSR) which intuitively measures the impact of shuffling on the decay rate of the system's state. We show how this quantity relates to stability properties of such systems. Specifically, from the SJSR, we can build a lower bound on the minimal shuffling rate in order to stabilize an unstable system. Then, we present several methods to approximate the SJSR, mainly by computing lower and upper bounds using Lyapunov methods and some automata theoretic techniques.

Keywords: Switched Systems, Lyapunov Functions, Exponential Stability, Formal Languages and Automata.

1 INTRODUCTION

Switched systems are dynamical systems with several modes of operations where the active mode is determined by a switching signal. Early works focus on proving stability of switched systems driven by arbitrary switching signals or by switching signals with dwell-time conditions [Liberzon (2003); Sun (2006); Lin and Antsaklis (2009)]. More recent works have considered systems with constrained switching signals where the switching signals are generated by labeled graphs [Lee and Dullerud (2007); Dai (2012); Athanasopoulos and Lazar (2014); Philippe et al. (2016); Pepe (2018)]. Shuffled switching signals is also a class of constrained switching signals that has been considered in the literature [Gurvits (1995); Wang et al. (2017); Girard and Mason (2019)]. A switching signal is said to be shuffled if all the modes of the switched systems are activated infinitely often. In this paper, we focus on the discrete-time switched linear systems driven by shuffled switching signals.

It is well known that the question of stability of discrete-time switched systems with arbitrary switching can be answered efficiently with the notion of the joint spectral radius (JSR) (see Jungers (2009) and the references therein). Intuitively the JSR represents the maximal asymptotic growth rate of products of matrices taken in a set. Computing the JSR is not an easy task but several techniques exist to approximate it. With a similar approach, it is possible to define a joint spectral radius related to the shuffled switched systems: the shuffled joint spectral radius (SJSR).

In this talk, we will present the notion of shuffled switched systems, we will also introduce the SJSR and investigate its properties. Then we will relate this quantity with the stability of such systems and finally we will describe numerical techniques for the SJSR approximation. Also, we mention that the proofs of theoretical results and numerical examples can be found in the preprint [Aazan et al. (2021)].

2 SHUFFLED JOINT SPECTRAL RADIUS

In this section, we present the concept of shuffled switched systems, then we introduce the notion of the ρ-SJSR, also we state a theorem that relates this quantity with the shuffled switched systems trajectories. Finally we present a resulting corollary that relates the system stabilizability with the ρ-SJSR.

2.1 Definition

We consider a discrete-time switched linear system described by the following equation

$$x(t+1) = A_{\theta(t)}x(t), \tag{1}$$

where $t \in \mathbb{N}$, $x(t) \in \mathbb{R}^n$ is the state and $\theta : \mathbb{N} \to \mathcal{I}$ is the switching signal belonging to a particular class of arbitrary switching signals: the shuffled switching signals. $\mathcal{I} = \{1, \cdots, m\}$, with $m \geq 2$ is the finite set of *modes* and $\mathcal{A} = \{A_i \in \mathbb{R}^{n\times n} | i \in \mathcal{I}\}$ is a finite set of matrices indexed by the modes. For a switching signal θ, let $\mathbb{A}_{\theta,0} = I_n$, and

⋆ This work was supported in part by the Agence Nationale de la Recherche (ANR) under Grant HANDY ANR-18-CE40- 0010.

$$\mathbb{A}_{\theta,T} = \prod_{t=0}^{T-1} A_{\theta(t)} = A_{\theta(T-1)} \cdots A_{\theta(0)},\ \forall T \geq 1.$$

Given an initial state $x_0 \in \mathbb{R}^n$, the trajectory defined by (1) with $x(0) = x_0$ is denoted $\mathbf{x}(.,x_0,\theta)$, it satisfies for all $t \in \mathbb{N}$, $\mathbf{x}(t,x_0,\theta) = \mathbb{A}_{\theta,t}x_0$.

Formally, the shuffled switching signal is defined as in [Girard and Mason (2019)]:

Definition 1. A switching signal $\theta : \mathbb{N} \to \mathcal{I}$ is *shuffled* if

$$\forall i \in \mathcal{I}, \forall T \in \mathbb{N}, \exists t \geq T :\ \theta(t) = i.$$

Following the previous definition, it is natural to define the following quantities related to a shuffled switching signal θ:

- The sequence of *shuffling instants* $(\tau_k^\theta)_{k\in\mathbb{N}}$ is defined by $\tau_0^\theta = 0$ and for all $k \in \mathbb{N}$,
$$\tau_{k+1}^\theta = \min\left\{ t > \tau_k^\theta \;\middle|\; \begin{array}{c} \forall i \in \mathcal{I}, \exists s \in \mathbb{N} : \\ \tau_k^\theta \leq s < t \text{ and } \theta(s) = i \end{array} \right\}.$$
- The *shuffling index* $\kappa^\theta : \mathbb{N} \to \mathbb{N}$ is given by
$$\kappa^\theta(t) = \max\{k \in \mathbb{N} |\ \tau_k^\theta \leq t\}.$$
- The *shuffling rate* γ^θ is defined as
$$\gamma^\theta = \liminf_{t\to+\infty} \frac{\kappa^\theta(t)}{t}.$$

Let $\mathcal{S}_s(\mathcal{I})$ be the set of all shuffled switching signals taking values in $\mathcal{I}$.

Now we are in a good position to define the ρ-SJSR. Given a finite set of matrices $\mathcal{A} \subseteq \mathbb{R}^{n\times n}$, let $\boldsymbol{\rho}(\mathcal{A})$ be its joint spectral radius (JSR), we recall that the JSR of a set of matrices $\mathcal{A} \subseteq \mathbb{R}^{n\times n}$ is defined as following:

$$\boldsymbol{\rho}(\mathcal{A}) = \lim_{k\to+\infty} \left(\sup \left\{ \left\| \prod_{j=1}^{k} A_j \right\|^{1/k} \;\middle|\; \begin{array}{c} A_j \in \mathcal{A}, \\ j = 1,\dots,k \end{array} \right\} \right).$$

We define the ρ-SJSR as following:

Definition 2. For all $\rho > \boldsymbol{\rho}(\mathcal{A})$, the Shuffled Joint Spectral Radius relative to $(\mathcal{A},\rho)$ (ρ-SJSR for short) is defined as

$$\boldsymbol{\lambda}(\mathcal{A},\rho) = \limsup_{k\to+\infty} \left(\sup_{\theta\in\mathcal{S}_s(\mathcal{I})} \left(\frac{\left\|\mathbb{A}_{\theta,\tau_k^\theta}\right\|}{\rho^{\tau_k^\theta}} \right)^{1/k} \right). \quad (2)$$

It is useful to say that the *limsup* in the above definition can be replaced by a simple limit. Due to the norm equivalence, one can replace the norm in (2) by any submultiplicative matrix norm.

2.2 Shuffled switched systems and ρ-SJSR

Now, when $\boldsymbol{\lambda}(\mathcal{A},\rho) > 0$, we bring out the relation between the system's trajectories and the ρ-SJSR. Then, we will derive a sufficient condition for stabilization based on the minimal shuffling rate and the ρ-SJSR. The following theorem clarifies the relationship between the ρ-SJSR and the behavior of the trajectories of (1).

Theorem 1. For all $\rho > \boldsymbol{\rho}(\mathcal{A})$, for all $\lambda \in (\boldsymbol{\lambda}(\mathcal{A},\rho), 1]$, there exists $C \geq 1$ such that

$$\begin{aligned} \|\mathbf{x}(t,x_0,\theta)\| \leq & C\rho^t\lambda^{\kappa^\theta(t)}\|x_0\|, \\ & \forall\theta \in \mathcal{S}_s(\mathcal{I}),\ \forall x_0 \in \mathbb{R}^n,\ \forall t \in \mathbb{N}. \end{aligned} \quad (3)$$

Conversely, if there exists $C \geq 1$, $\rho \geq 0$ and $\lambda \in [0,1]$ such that (3) holds, then either $\rho > \boldsymbol{\rho}(\mathcal{A})$ and $\lambda \geq \boldsymbol{\lambda}(\mathcal{A},\rho)$, or $\rho = \boldsymbol{\rho}(\mathcal{A})$ and $\lambda \geq \sup_{\rho'>\boldsymbol{\rho}(\mathcal{A})} \boldsymbol{\lambda}(\mathcal{A},\rho')$.

A remarkable result from the previous theorem is a sufficient condition for stabilization based on the minimal shuffling rate.

Corollary 1. Assume $\boldsymbol{\lambda}(\mathcal{A},\rho) > 0$ for every $\rho > \boldsymbol{\rho}(\mathcal{A})$. Let $\theta \in \mathcal{S}_s(\mathcal{I})$, if there exists $\rho > \boldsymbol{\rho}(\mathcal{A})$ such that $\gamma^\theta > -\frac{\ln(\rho)}{\ln(\boldsymbol{\lambda}(\mathcal{A},\rho))}$, then

$$\lim_{t\to+\infty} \|\mathbf{x}(t,x_0,\theta)\| = 0, \quad \forall x_0 \in \mathbb{R}^n. \quad (4)$$

The proof of this corollary follows from the previous theorem and from the shuffling rate definition.

3 APPROXIMATION OF THE ρ-SJSR

We have seen in the previous section that the ρ-SJSR plays an important role in the stability analysis of shuffled switched systems. However, like the JSR, this quantity is difficult to calculate. In this section, we will consider the problem of approximating the ρ-SJSR. In the following, we will give an explicit expression for a lower bound based on the JSR of a set constructed from $\mathcal{A}$, we will show that this lower bound is asymptotically tight, moreover an exact expression for the ρ-SJSR will be given under certain conditions. Next, an approach to find upper bounds will be given based on multiple Lyapunov functions and automata theoretic techniques.

3.1 Lower bounds computation

Let $\mathcal{N}_\mathcal{I}$ be the set of products of matrices where all modes in $\mathcal{I}$ appear exactly once, formally:

$$\mathcal{N}_\mathcal{I} = \left\{ \prod_{k=1}^{m} A_{j_k} \;\middle|\; \begin{array}{c} j_1,\dots,j_m \in \mathcal{I}, \\ \forall i \in \mathcal{I}, \exists k \in \{1,\dots,m\}, j_k = i \end{array} \right\}.$$

The following theorem gives an explicit expression of a lower bound for the ρ-SJSR based on the JSR of $\mathcal{N}_\mathcal{I}$. Also, it shows that this lower bound is asymptotically tight, moreover, under certain conditions, it reveals an explicit expression of the ρ-SJSR.

Theorem 2. The following results hold true.

(i) For all $\rho > \boldsymbol{\rho}(\mathcal{A})$,
$$\boldsymbol{\lambda}(\mathcal{A},\rho) \geq \frac{\boldsymbol{\rho}(\mathcal{N}_\mathcal{I})}{\rho^m}. \quad (5)$$

(ii) We have the asymptotic estimate
$$\lim_{\rho\to+\infty} \rho^m \boldsymbol{\lambda}(\mathcal{A},\rho) = \boldsymbol{\rho}(\mathcal{N}_\mathcal{I}). \quad (6)$$

(iii) If there exists a norm $\|\cdot\|_*$ that is extremal [1] for $\mathcal{N}_\mathcal{I}$, then there exists $R \geq \boldsymbol{\rho}(\mathcal{A})$ such that for all $\rho \geq R$,
$$\boldsymbol{\lambda}(\mathcal{A},\rho) = \frac{\boldsymbol{\rho}(\mathcal{N}_\mathcal{I})}{\rho^m}. \quad (7)$$

Remark: If $\mathcal{A}$ consists of invertible matrices only, an explicit expression of R can be given.

In the next section, we give a method for computing upper bounds using Lyapunov theory.

[1] An induced norm $\|.\|_*$ is said to be extremal for a set of matrices $\mathcal{A}$, if it satisfies $\boldsymbol{\rho}(\mathcal{A}) = \max_{A\in\mathcal{A}} \|A\|_*$.

3.2 Upper bounds computation

This section details a method for computing upper bounds on the JSR and the ρ-SJSR.

Theorem 3. If there exist $V : (2^{\mathcal{I}} \setminus \{\mathcal{I}\}) \times \mathbb{R}^n \to \mathbb{R}_0^+$, $\alpha_1, \alpha_2, \rho > 0$ and $\lambda \in [0, 1]$ such that the following inequalities hold true for every $x \in \mathbb{R}^n$

$$\alpha_1 \|x\|^2 \leq V(J, x) \leq \alpha_2 \|x\|^2, \qquad \forall J \subsetneq \mathcal{I} \tag{8}$$

$$V(J \cup \{i\}, A_i x) \leq \rho^2 V(J, x), \qquad \text{if } J \cup \{i\} \neq \mathcal{I} \tag{9}$$

$$V(\emptyset, A_i x) \leq \rho^2 \lambda^2 V(J, x), \qquad \text{if } J \cup \{i\} = \mathcal{I} \tag{10}$$

then the bound (3) holds. Conversely, if the matrices A_i are invertible, for all $i \in \mathcal{I}$ and the bound (3) holds for some $\rho > 0$, $\lambda \in [0, 1]$ and $C \geq 1$, then there exists a function $V : (2^{\mathcal{I}} \setminus \{\mathcal{I}\}) \times \mathbb{R}^n \to \mathbb{R}_0^+$ such that the inequalities (8), (9) and (10) are satisfied.

The proof of the direct result relies on a finite state automaton whose states corresponds to subsets of $\mathcal{I}$, where each time the switching signal shuffles, there will be a transition to the automaton's initial state which corresponds to the empty set, therefore, based on the automaton, it is not hard to construct the conditions of the theorem. The proof of the converse result relies on a multiple Lyapunov function (each corresponds to a state) which can be seen as the supremum of trajectories norm that lead from the corresponding state to a specific state.

By combining the result of the previous theorem with Theorem 1, one can easily compute upper bounds on the ρ-SJSR and the JSR using some LMIs.

4 CONCLUSION

In this work, we have defined the ρ-SJSR, a special kind of JSR related to shuffled switched systems. We successfully related this notion to the stabilization of such systems. Also, we provided some theoretical tools for approximation using automata theoretic techniques and Lyapunov functions. Some interesting numerical examples can be found in [Aazan et al. (2021)] and will be presented in the talk.

The current work opens several research directions for the future. First, the development of numerical and theoretic techniques to compute tighter bounds on the ρ-SJSR. Secondly since our approach is based on automata theoretic techniques, it is natural to think that one can derive stability conditions by working directly on the Büchi, Rabin or Muller automaton specifying the ω-regular language.

References

Aazan, G., Girard, A., Greco, L., and Mason, P. (2021). Stability of shuffled switched linear systems: A joint spectral radius approach. URL `https://hal.archives-ouvertes.fr/hal-03257026`.

Athanasopoulos, N. and Lazar, M. (2014). Stability analysis of switched linear systems defined by graphs. In *IEEE Conference on Decision and Control*, 5451–5456.

Berger, M.A. and Wang, Y. (1992). Bounded semigroups of matrices. *Linear Algebra and its Applications*, 166, 21–27.

Dai, X. (2012). A Gel'fand-type spectral radius formula and stability of linear constrained switching systems. *Linear Algebra and its Applications*, 436(5), 1099–1113.

Girard, A. and Mason, P. (2019). Lyapunov functions for shuffle asymptotic stability of discrete-time switched systems. *IEEE Control Systems Letters*, 3(3), 499–504.

Gurvits, L. (1995). Stability of discrete linear inclusion. *Linear algebra and its applications*, 231, 47–85.

Jungers, R. (2009). *The joint spectral radius: theory and applications*, volume 385. Springer Science & Business Media.

Lee, J.W. and Dullerud, G.E. (2007). Uniformly stabilizing sets of switching sequences for switched linear systems. *IEEE Transactions on Automatic Control*, 52(5), 868–874.

Liberzon, D. (2003). *Switching in systems and control*. Springer Science & Business Media.

Lin, H. and Antsaklis, P.J. (2009). Stability and stabilizability of switched linear systems: a survey of recent results. *IEEE Transactions on Automatic control*, 54(2), 308–322.

Pepe, P. (2018). Converse Lyapunov theorems for discrete-time switching systems with given switches digraphs. *IEEE Transactions on Automatic Control*.

Philippe, M., Essick, R., Dullerud, G.E., and Jungers, R.M. (2016). Stability of discrete-time switching systems with constrained switching sequences. *Automatica*, 72, 242–250.

Sun, Z. (2006). *Switched linear systems: control and design*. Springer Science & Business Media.

Wang, Y., Roohi, N., Dullerud, G.E., and Viswanathan, M. (2017). Stability analysis of switched linear systems defined by regular languages. *IEEE Transactions on Automatic Control*, 62(5), 2568–2575.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A time-varying approach for model reduction of singular linear switched systems in discrete time

Md. Sumon Hossain * **Sutrisno** *,** **Stephan Trenn** *

* *Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen, The Netherlands.*
** *Dept. of Mathematics, Diponegoro University, Indonesia.*

Abstract: We propose a model reduction approach for singular linear switched systems in discrete time with a fixed mode sequence based on a balanced truncation reduction method for linear time-varying discrete-time systems. The key idea is to use the one-step map to find an equivalent time-varying system with an identical input-output behavior, and then adapt available balance truncation methods for (discrete) time-varying systems. The proposed method is illustrated with a low-dimensional academic example.

Keywords: singular linear switched systems, time-varying systems, reachability and observability Gramians and balanced truncation.

1. INTRODUCTION

In this paper we consider singular linear switched systems (SLSSs) in discrete time of the form

$$S_\sigma : \begin{cases} E_{\sigma(k)}x(k+1) = A_{\sigma(k)}x(k) + B_{\sigma(k)}u(k), \\ y(k) = C_{\sigma(k)}x(k), \ k \in \mathbb{N}, \end{cases} \tag{1}$$

where $x(k) \in \mathbb{R}^n$ is the state at time $k \in \mathbb{N}$ and $\sigma : \mathbb{N} \to Q = \{0, 1, 2, \cdots, \mathtt{m}\}$, $\mathtt{m} \in \mathbb{N}$, is the switching signal with the switching times $0 < s_1 < s_2 < \cdots < s_\mathtt{m}$ in the bounded interval $[k_0, k_f) := \{k_0, k_0 + 1, \ldots, k_f - 1\}$ of interest. The system matrices are $E_i, A_i \in \mathbb{R}^{n\times n}$, $B_i \in \mathbb{R}^{n\times m}$, $C_i \in \mathbb{R}^{p\times n}$, where $i \in Q$. The matrices E_i are in general singular, which is related to the presence of (mode-dependent) algebraic constraints. We assume that the i-th mode is active in the interval $[s_i, s_{i+1})$, for $i = 0, 1, \cdots, \mathtt{m}$ (where $s_0 := 0$) and define the duration of the i-th mode as $\tau_i = s_{i+1} - s_i$. Since we will be interested in the input-output behavior of S_σ we assume in the following that $x(0) = 0$.

Control problems governed by SLSSs arise in a variety of practical applications including circuit simulation, computational electromagnetics, fluid dynamics, mechanical and chemical engineering; see Luenberger (1977); Xia et al. (2008). In some cases, these systems lead to analyzing large-scale and complex dynamical systems. Although, the computational speed and performance of the modern computers are increasing; simulation, optimization or real time controller design for such large-scale systems are still difficult due to extra memory requirements and additional computational complexity. Model order reduction (MOR) is a useful tool for dealing with such complexity, wherein one seeks a simpler model that can then be used as an efficient surrogate model to the original model. There are already some existing results on MOR for switched ODEs, see e.g. Schulze and Unger (2018); Gosea et al. (2020) for continuous time case, and Baştuğ et al. (2016, 2014); Shaker and Wisniewski (2012); Birouche et al. (2012) for discrete time case. However, in contrast to the existing literature, we view here the SLSS (1) as a *time-varying* linear systems, in particular, the reduction in general depends on the specifically given switching signal and results in a time-varying reduced model.

The remaining paper is structured as follows. We discuss the problem formulation and some preliminaries for singular system in Section 2. Section 3 provides the computation procedure of time-varying balanced realization in discrete time. In Section 4, we present time-varying balanced truncation method for SLSS. Finally, some numerical results are presented in Section 5.

2. PRELIMINARIES AND PROBLEM STATEMENT

In this section, it is shown that the solutions of a SLSS can equivalently be expressed in terms of a time-varying system. For the existence and uniqueness of solutions of SLSSs the following assumption is needed.

Assumption 1. The SLSS (1) is *jointly index-1*, i.e.

$$\mathcal{S}_i \oplus \ker E_j = \mathbb{R}^n, \quad \forall i, j \in Q,$$

where $\mathcal{S}_i = A_i^{-1}(\operatorname{im} E_i)$. ◇

Under the jointly index-1 assumption, the solution of SLSS (1) with $x(0) = 0$ exists. This solution is unique and satisfies the following lemma.

Lemma 1. (Cf. Anh et al. (2019)) Assume the SLSS (1) is jointly index-1. For a given switching signal σ, there exist corresponding matrices $\widetilde{A}_k$, $\widetilde{B}_k$ and $\widetilde{F}_k$, such that all solutions of (1) with $x(0) = 0$ satisfy

$$x(k+1) = \widetilde{A}_k x(k) + \widetilde{B}_k u(k) + \widetilde{F}_k u(k+1), \quad k \in \mathbb{N}. \tag{2}$$

⋆ Email Addresses: s.hossain@rug.nl; s.sutrisno[@rug.nl, @live.undip.ac.id]; s.trenn@rug.nl;

Proof. Let $\sigma(-1) := \sigma(0)$ and, for $k \in \mathbb{N}$,

$$\widetilde{A}_k := V_{\sigma(k)} \begin{bmatrix} \bar{A}^1_{\sigma(k),\sigma(k-1)} & 0 \\ -\bar{A}^2_{\sigma(k+1),\sigma(k)}\bar{A}^1_{\sigma(k),\sigma(k-1)} & 0 \end{bmatrix} V^{-1}_{\sigma(k-1)}, \quad (3a)$$

$$\widetilde{B}_k := V_{\sigma(k)} \begin{bmatrix} \bar{B}^1_{\sigma(k),\sigma(k-1)} \\ -\bar{A}^2_{\sigma(k+1),\sigma(k)}\bar{B}^1_{\sigma(k),\sigma(k-1)} \end{bmatrix}, \quad (3b)$$

$$\widetilde{F}_k := V_{\sigma(k)} \begin{bmatrix} 0 \\ -\bar{B}^2_{\sigma(k+1),\sigma(k)} \end{bmatrix}, \quad (3c)$$

where

$$\begin{bmatrix} \bar{A}^1_{i,j} & 0 \\ -\bar{A}^2_{i,j} & I_{n_2} \end{bmatrix} = V_i^{-1} G_{i,j}^{-1} A_i V_j, \quad \begin{bmatrix} \bar{B}^1_{i,j} \\ \bar{B}^2_{i,j} \end{bmatrix} = V_i^{-1} G_{i,j}^{-1} B_i,$$

$$G_{i,j} = E_i + A_i Q_{i,j}, \quad Q_{i,j} = V_j \begin{bmatrix} 0 & 0 \\ 0 & I_{n_2} \end{bmatrix} V_i^{-1},$$

$V_i = [g_i^1, ..., g_i^{n_1}, h_i^{n_1+1}, ..., h_i^n]$, $g_i^1, ..., g_i^{n_1}$ are the bases of $\mathcal{S}_i$, and $h_i^{n_1+1}, ..., h_i^n$ are the bases of $\ker E_i$. The remaining proof is similar to the proof of (Anh et al., 2019, Thm. 5.1) and therefore omitted. ■

Remark 2. The one-step map from $x(k)$ to $x(k+1)$ depends on the modes at time $k-1$, k and $k+1$. This concludes that the allowed space of consistent initial values also depends on the choice of $\sigma(-1)$, here we assume that $\sigma(-1) = \sigma(0)$. As pointed out in (Anh et al., 2019, Rem. 5.2), the effect of a different choice $\sigma(-1)$ is not yet fully understood, and is still under investigation; nevertheless, since we restrict our attention to the initial condition $x(0) = 0$, this is of no further concern to us here.

Motivated by Lemma 1, we consider the following time-varying surrogate system for (1) with given switching signal σ:

$$\widetilde{S}_\sigma : \begin{cases} x(k+1) = \widetilde{A}_k x(k) + \begin{bmatrix} \widetilde{B}_k & \widetilde{F}_k \end{bmatrix} \widetilde{u}(k), \\ y(k) = C_k x(k), \ k \in \mathbb{N}, \end{cases} \quad (4)$$

where $x(0) = 0$, $\widetilde{u}(k) = \begin{bmatrix} u(k) \\ u(k+1) \end{bmatrix}$, $C_k := C_{\sigma(k)}$ and $\widetilde{A}_k$, $\widetilde{B}_k$, $\widetilde{F}_k$ are given by (3). Writing $\tilde{u} = \begin{bmatrix} I \\ \mathcal{T}_1 \end{bmatrix} u$, where $\mathcal{T}_1\{u\}(k) := u(k+1)$ denotes the time-shift operator, by Lemma 1, (1) and (4) have the same input-output behaviour.

Note that the solution of jointly index-1 SLSS (1) does not exist for any initial value $x(0) \in \mathbb{R}^n$. In fact, the consistency space of jointly index-1 (1), under the assumption $\sigma(-1) = \sigma(0)$, is $\operatorname{im} V_{\sigma(0)} \begin{bmatrix} I & 0 \\ -\widehat{A}^2_{\sigma(0),\sigma(0)} & \widehat{B}^2_{\sigma(0),\sigma(0)} \end{bmatrix}$. This has some implications on the relationship between system S_σ and $\widetilde{S}_\sigma$ in terms of observability and reachability. Here, observability means that if the input and output are identically zero on $[k_0, k_f)$ also the state has to be zero; reachability means, that for each $x_f \in \mathbb{R}^n$, there exists an input such that the corresponding solution satisfies $x(k_f - 1) = x_f$. Clearly, a reachable SLSS S_σ implies a reachable time varying surrogate system $\widetilde{S}_\sigma$ whereas an observable SLSS S_σ does not imply that its surrogate system $\widetilde{S}_\sigma$ is observable. However, an observable $\widetilde{S}_\sigma$ implies that S_σ is also observable.

Our goal is to find for the time-varying system (4) a reduced size time-varying system

$$\widehat{S}_\sigma : \begin{cases} \widehat{x}(k+1) = \widehat{A}_k \widehat{x}(k) + \begin{bmatrix} \widehat{B}_k & \widehat{F}_k \end{bmatrix} \begin{bmatrix} u(k) \\ u(k+1) \end{bmatrix}, \\ \widehat{y}(k) = \widehat{C}_k \widehat{x}(k), \ k \in \mathbb{N}. \end{cases} \quad (5)$$

with reduced system matrices $\widehat{A}_i \in \mathbb{R}^{r \times r}$, $\widehat{B}_i, \widehat{F}_i \in \mathbb{R}^{r \times m}$, $\widehat{C}_i \in \mathbb{R}^{p \times r}$ and $r \ll n$, such that $\widehat{y} \approx y$ for a large class of inputs u. Due to the input-output equivalence between (1) and (4), the reduced system (5) will then also be good surrogate model for the original SLSS.

3. TIME-VARYING BALANCED REALIZATIONS

3.1 Time-varying Gramians

Consider a time-varying discrete time system of the form

$$\begin{aligned} x(k+1) &= A_k x(k) + B_k u(k), \quad k \in [k_0, k_f) \\ y(k) &= C_k x(k). \end{aligned} \quad (6)$$

Definition 3. The time-varying reachability and observability Gramians of (6) are defined recursively as

$$P(k) = A_{k-1} P(k-1) A_{k-1}^\top + B_{k-1} B_{k-1}^\top, \quad (7)$$

$$Q(k) = A_k^\top Q(k+1) A_k + C_k^\top C_k, \quad (8)$$

with some positive semi-definite initial/final values $P(k_0) = P_0$ and $Q(k_f) = Q_f$

Note that the reachability Gramian is constructed *forward* in time, while the observability Gramians evolves *backward* in time.

Remark 4. The choice of the initial/final Gramians is crucial in the sense that they play an important role for the magnitude of all other subsequent Gramians. At this moment the best choice of the initial/final Gramians is not clear. In the context of time-varying case, two versions can be proposed for the initial/final Gramians. One choice could be to assume that the first mode is active in the past, i.e. $(-\infty, k_0]$, and the Gramians of the first mode is considered as the initial reachability Gramian. Similarly, by assuming that the last mode will be active in the future, i.e. $[k_f, \infty)$, and the Gramian of the last is considered as the final value for observability Gramian. However, in this choice, the computation of infinite Gramians is only possible for stable modes; here, we do not assume stability of each mode. On the other hand, a second choice could be the identity matrix which would not affect the direction of the states which are difficult to control and difficult to observe. By scaling the identity matrix with a smaller magnitude, one can restrict the influence of these artificial initial/final Gramians relatively to the time-varying Gramians and also for the bounded time-varying coordinate transformation matrices.

Note that, $P(k)$ and $Q(k)$ are both symmetric and positive semidefinite for all $k \in [k_0, k_f]$ if P_0 and Q_f are positive definite. It is assumed that the input-output balancing with respect to the reachability and observability Gramians is defined over specific time intervals. Hence, no assumption is needed with regard to the stability of the system.

Applying any time-varying coordinate transformation

$$x(k) = T(k)\overline{x}(k)$$

to (6) results in an equivalent system

$$\overline{x}(k+1) = \overline{A}_k\overline{x}(k) + \overline{B}_k u(k)$$
$$y(k) = \overline{C}_k x(k),$$

with $\overline{A}_k := T(k+1)^{-1}A_kT(k)$, $\overline{B}_k := T(k+1)^{-1}B_k$, $\overline{C}_k := C_kT(k)$. It is easily seen, the corresponding Gramians satisfy

$$\overline{P}(k) = T(k)^{-1}P(k)T(k)^{-\top},$$
$$\overline{Q}(k) = T(k)^{\top}Q(k)T(k),$$

if the initial/final values satisfy $\overline{P}_0 = T(0)^{-1}P_0T(0)^{-\top}$ and $\overline{Q}_f = T(k_f)^{\top}Q_fT(k_f)$. In particular,

$$\overline{P}(k)\overline{Q}(k) = T(k)^{-1}P(k)Q(k)T(k).$$

This shows that, under such transformation, the eigenvalues of the product of Gramians are invariant.

The key idea of balanced truncation is to find a coordinate transformation such that the corresponding Gramians become equal and diagonal. How to achieve such a balancing transformation is given in the following lemma.

Lemma 5. Assume that Gramiens $P(k)$ and $Q(k)$ of the time-varying system (6) are nonsingular on $[k_0, k_f)$. Then, there exists a transformation $T : [k_0, k_f] \to \mathbb{R}^{n\times n}$ such that

$$T(k)^{-1}P(k)T(k)^{-\top} = T(k)^{\top}Q(k)T(k) = \Xi(k), \qquad (9)$$

for all $k \in [k_0, k_f]$ and $\Xi(k) = \{\xi_1(k), \ldots, \xi_n(k)\}$ is a diagonal matrix. In fact, the transformation matrices are given by

$$T(k) = R(k)U(k)\Xi(k)^{-1/2},$$
$$T(k)^{-1} = \Xi(k)^{-1/2}V(k)^{\top}L(k)^{\top},$$

where $U(k)\Xi(k)V(k)^{\top}$ is the singular value decomposition of $R(k)^{\top}L(k)$, and where $R(k)R(k)^{\top} = P(k)$ and $L(k)L(k)^{\top} = Q(k)$ are the Cholesky decompositions of P and Q, respectively.

Proof. The proof is similar to the proof of (Hossain and Trenn, 2020, Lemma 11) and therefore omitted. ■

4. MODEL REDUCTION

We now combine the above results to propose a model reduction method for SLSS (1) based on balanced truncation. By Assumption 1, we can instead consider system (4) and we can construct the corresponding time-varying reachability/observability Gramians $\widetilde{P}(k)$ and $\widetilde{Q}(k)$ for $(\widetilde{A}_k, [\widetilde{B}_k, \widetilde{F}_k], \widetilde{C}_k)$ for some initial/final Gramians $\widetilde{P}_0, \widetilde{Q}_f$. Now an assumption is needed for model reduction methods.

Assumption 2. Assume a transformation $\widetilde{T}$ such that the balanced Gramians are obtained by

$$\widetilde{T}(k)^{-1}\widetilde{P}(k)\widetilde{T}(k)^{-\top} = \widetilde{T}(k)^{\top}\widetilde{Q}(k)\widetilde{T}(k) = \widetilde{\Xi}(k)$$

and let, the (uniformly) partitioned form $\widetilde{\Xi}(k) = \begin{bmatrix}\widehat{\Xi}(k) & 0\\ 0 & \overline{\Xi}(k)\end{bmatrix}$, where all diagonal entries in $\overline{\Xi}(k)$ are significantly smaller than those in $\widehat{\Xi}(k)$. ◇

With the Assumption 2, the singular value decomposition is then given by

$$\widetilde{R}(k)^{\top}\widetilde{L}(k) = \begin{bmatrix}\widehat{U}(k) & \overline{U}(k)\end{bmatrix}\begin{bmatrix}\widehat{\Xi}(k) & 0\\ 0 & \overline{\Xi}(k)\end{bmatrix}\begin{bmatrix}\widehat{V}(k) & \overline{V}(k)\end{bmatrix}^{\top}$$

where $\widetilde{R}(k)\widetilde{R}(k)^{\top} = \widetilde{P}(k)$ and $\widetilde{L}(k)\widetilde{L}(k)^{\top} = \widetilde{Q}(k)$ are obtained by a Cholesky decomposition. According to this splitting, let $\widetilde{T}(k) = [\widehat{\Pi}_R(k), *]$ and $\widetilde{T}(k)^{-1} = [\widehat{\Pi}_L(k), *]^{\top}$, and define

$$\widehat{A}_k := \widehat{\Pi}_L(k+1)\widetilde{A}_k\widehat{\Pi}_R(k),$$
$$\begin{bmatrix}\widehat{B}_k & \widehat{F}_k\end{bmatrix} := \widehat{\Pi}_L(k+1)\begin{bmatrix}\widetilde{B}_k & \widetilde{F}_k\end{bmatrix},$$
$$\widehat{C}_k := \widetilde{C}_k\widehat{\Pi}_R(k),$$

which results in our proposed reduced system (5), where the left- and right-projectors are calculated as

$$\widehat{\Pi}_R(k) := \widetilde{R}(k)\widehat{U}(k)\widehat{\Xi}(k)^{-1/2} \in \mathbb{R}^{n\times r},$$
$$\widehat{\Pi}_L(k) := \widehat{\Xi}(k)^{-1/2}\widehat{V}(k)^{\top}\widetilde{L}(k)^{\top} \in \mathbb{R}^{r\times n}.$$

5. NUMERICAL RESULTS

This section illustrates the proposed method by providing an example.

Example 6. Consider a SLSS with two modes

$$(E_0, A_0, B_0, C_0) = \left(\begin{bmatrix}1&0&0&0\\1&1&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}, \begin{bmatrix}1&0&0&0\\1&0&1&0\\1&1&1&0\\0&0&0&1\end{bmatrix}, \begin{bmatrix}0.02\\2\\1\\0.2\end{bmatrix}, \begin{bmatrix}-0.1\\0.1\\0.1\\2\end{bmatrix}^{\top}\right),$$
$$(E_1, A_1, B_1, C_1) = \left(\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}, \begin{bmatrix}1&0&0&0\\0&1&1&0\\1&0&1&0\\0&0&0&1\end{bmatrix}, \begin{bmatrix}0.01\\2\\0.5\\0.1\end{bmatrix}, C_0\right),$$

Consider a switching signal $\sigma : [0, 9) \to \{0, 1\}$,

$$\sigma(k) = \begin{cases}0 & : k \in [0, 4) \cup [7, 9),\\ 1 & : k \in [4, 7).\end{cases}$$

It can easily be verified that the pairs (E_0, A_0) and (E_1, A_1) are jointly index-1. Hence, by Lemma 1, the time-varying system (4) is obtained with the following system matrices

$$(\widetilde{A}_k, \widetilde{B}_k) = \begin{cases}\left(\begin{bmatrix}1&0&0&0\\0&0&1&0\\1&1&1&0\\0&0&0&0\end{bmatrix}, \begin{bmatrix}0.02\\1.98\\1\\0\end{bmatrix}\right) & : k = 0, 1, 2, 3, 7, 8,\\ \left(\begin{bmatrix}1&0&0&0\\0&1&1&0\\1&0&1&0\\0&0&0&0\end{bmatrix}, \begin{bmatrix}0.01\\2\\0.5\\0\end{bmatrix}\right) & : k = 4, 5, 6,\end{cases}$$
$$\widetilde{F}_k = \begin{cases}\begin{bmatrix}0\\0\\0\\-0.2\\0\end{bmatrix} & : k = 0, 1, 2, 6, 7, 8,\\ \begin{bmatrix}0\\0\\0\\-0.1\end{bmatrix} & : k = 3, 4, 5.\end{cases}$$

The corresponding reachability and observability Gramians are calculated respectively, $\widetilde{P}(k)$ and $\widetilde{Q}(k)$ for $k \in [0, 9)$ with initial/final values $\widetilde{P}_0 = 0.002I$ and $\widetilde{Q}_f = 0.002I$. The corresponding HSVs are depicted in Figure 1 and it is apparent that the last two HSVs are significantly smaller than the first two. Hence, a two dimensional reduced system is obtained which approximates the time-varying system (4) and hence, the original SLSS.

The computed two dimensional reduced systems at each time steps are given by $(\widehat{A}_k, [\widehat{B}_k, \widehat{F}_k], \widehat{C}_k) =$

$$\left(\begin{bmatrix}0.9206 & -0.0051\\ -0.0107 & 0.0012\end{bmatrix}\begin{bmatrix}-1.8615 & 0.0046\\ -0.0535 & 0.6305\end{bmatrix}\begin{bmatrix}-0.1410\\ -0.6334\end{bmatrix}^{\top}\right),$$
$$\left(\begin{bmatrix}0.9761 & -0.0071\\ -0.0058 & -0.0076\end{bmatrix}\begin{bmatrix}-1.0832 & 0.0074\\ -0.0603 & 0.6287\end{bmatrix}\begin{bmatrix}-0.2387\\ -0.6332\end{bmatrix}^{\top}\right),$$
$$\left(\begin{bmatrix}0.9887 & -0.0116\\ -0.0027 & -0.0071\end{bmatrix}\begin{bmatrix}-0.7265 & 0.0117\\ -0.0445 & 0.8861\end{bmatrix}\begin{bmatrix}-0.3859\\ -0.4449\end{bmatrix}^{\top}\right),$$

$$\left(\begin{bmatrix} 1.1336 & -0.0155 \\ -0.4762 & -0.0398 \end{bmatrix} \begin{bmatrix} -0.3934 & 0.0143 \\ -0.0311 & 0.4016 \end{bmatrix} \begin{bmatrix} -0.7193 \\ -0.4515 \end{bmatrix}^\top\right),$$
$$\left(\begin{bmatrix} 0.9856 & -0.0335 \\ -0.0019 & 0.1912 \end{bmatrix} \begin{bmatrix} -0.2249 & 0.0206 \\ -0.0536 & 0.4158 \end{bmatrix} \begin{bmatrix} -1.0114 \\ -0.4262 \end{bmatrix}^\top\right),$$
$$\left(\begin{bmatrix} 0.9719 & -0.0475 \\ -0.0036 & 0.0701 \end{bmatrix} \begin{bmatrix} -0.1850 & 0.0286 \\ -0.0047 & 0.2986 \end{bmatrix} \begin{bmatrix} -1.3323 \\ -0.5423 \end{bmatrix}^\top\right),$$
$$\left(\begin{bmatrix} 0.8014 & -0.0770 \\ -0.8119 & 0.0689 \end{bmatrix} \begin{bmatrix} -0.1249 & 0.0523 \\ 0.1196 & -0.5938 \end{bmatrix} \begin{bmatrix} -1.4777 \\ 0.5432 \end{bmatrix}^\top\right),$$
$$\left(\begin{bmatrix} 0.9471 & 0.0834 \\ -0.0047 & 0.0125 \end{bmatrix} \begin{bmatrix} -0.1026 & 0.1118 \\ -0.0261 & 0.4245 \end{bmatrix} \begin{bmatrix} -2.8391 \\ -0.1943 \end{bmatrix}^\top\right).$$

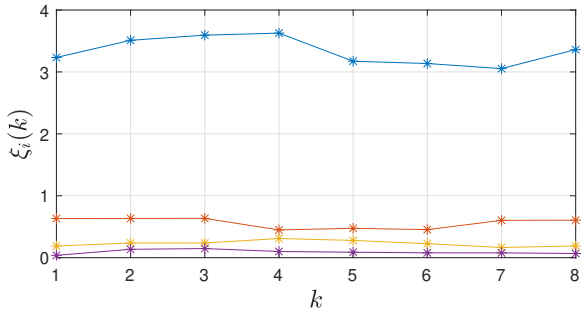


Fig. 1. Hankel singular values of balanced Gramians at each time instance.

Consider randomly generated input $u(\cdot)$ with $u(0) = 0$, and the input-output behavior is calculated for the system (4) and its reduced system with relative errors. Figure 2 displays the output, the input signal, and the relative error for the original system and the proposed two dimensional reduced system. Clearly, both outputs match nicely and the relative error is less then 6%.

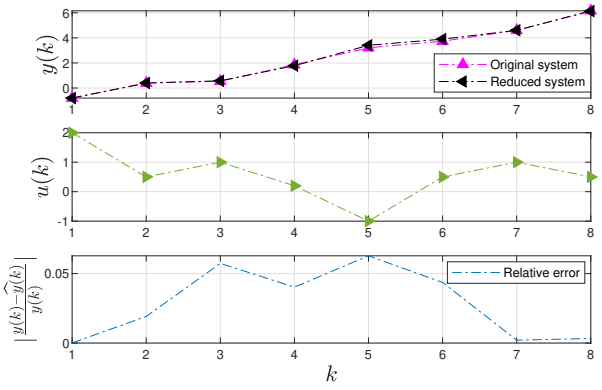


Fig. 2. Outputs and the relative error of the original system (4) and the proposed 2^{nd} order approximation.

Next, another initial /final value of the Gramians is considered by increasing the magnitudes as $\widetilde{P}_0 = 0.5I$ and $\widetilde{Q}_f = 0.5I$. With the same input sequence as in Figure 2, the input-output behavior with the relative error is depicted in Figure 3, which shows that the choice of the initial /final values of Gramians plays an important role in the error analysis. Therefore, it is concluded that taking small magnitude with identity matrix could be the best choice for the initial /final values of the Gramians.

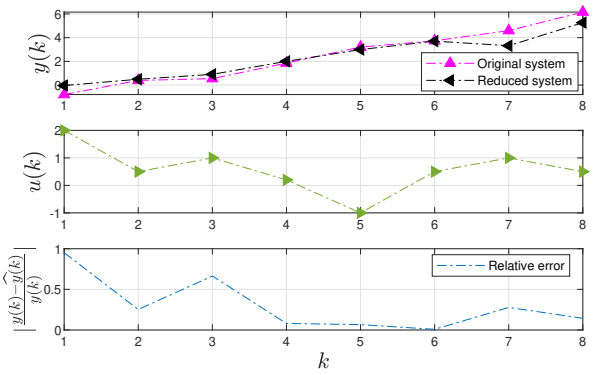


Fig. 3. Outputs and relative errors of the original system (4) and the proposed 2nd order approximation with initial/final values $\widetilde{P}_0 = 0.5I$, $\widetilde{Q}_f = 0.5I$.

6. CONCLUSION

In this paper, we have presented a time-varying approach for proposing a reduced system for singular linear switched systems. The key novelty is the viewpoint of the SLSS as a piecewise-constant time-varying system. At first, we have focused on input-extended time-varying ODEs, which gives identical input-output behavior as the original index-1 SLSSs. Then, by applying the well known time-varying balanced truncation method for the discrete time case, we find a good approximation of the time-varying system.

REFERENCES

Anh, P.K., Linh, P.T., Trenn, S., et al. (2019). The one-step-map for switched singular systems in discrete-time. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 605–610. IEEE.

Baştuğ, M., Petreczky, M., Wisniewski, R., and Leth, J. (2014). Model reduction of linear switched systems by restricting discrete dynamics. In *53rd IEEE Conference on Decision and Control*, 4422–4427. IEEE.

Baştuğ, M., Petreczky, M., Wisniewski, R., and Leth, J. (2016). Reachability and observability reduction for linear switched systems with constrained switching. *Automatica*, 74, 162–170.

Birouche, A., Mourllion, B., and Basset, M. (2012). Model order-reduction for discrete-time switched linear systems. *International Journal of Systems Science*, 43(9), 1753–1763.

Gosea, I.V., Duff, I.P., Benner, P., and Antoulas, A.C. (2020). Model order reduction of switched linear systems with constrained switching. In *IUTAM Symposium on Model Order Reduction of Coupled Systems, Stuttgart, Germany, May 22–25, 2018*, 41–53. Springer.

Hossain, M.S. and Trenn, S. (2020). A time-varying gramian based model reduction approach for linear switched systems. *IFAC-PapersOnLine*, 53(2), 5629–5634. 21th IFAC World Congress.

Luenberger, D.G. (1977). Dynamic equations in descriptor form. *IEEE Trans. Autom. Control*, 22, 312–321.

Schulze, P. and Unger, B. (2018). Model reduction for linear systems with low-rank switching. *SIAM Journal on Control and Optimization*, 56(6), 4365–4384.

Shaker, H.R. and Wisniewski, R. (2012). Model reduction of switched systems based on switching generalized gramians. *International Journal of Innovative Computing, Information and Control*, 8(7), 5025–5044.

Xia, Y., Zhang, J., and Boukas, E.K. (2008). Control for discrete singular hybrid systems. *Automatica*, 44(10), 2635–2641.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Genetic Algorithms with Permutation-Based Representation for Computing the Distance of Linear Codes ⋆

M. P. Cuéllar [**,***] **J. Gómez-Torrecillas** [*,****]
F. J. Lobillo [*,***] **G. Navarro** [**,***]

[*] *Department of Algebra, University of Granada*
[**] *Department of Computer Science and Artificial Intelligence, University of Granada, Spain*
[***] *CITIC, University of Granada*
[****] *IEMath-GR, University of Granada*

Abstract: Finding the minimum distance of linear codes is an NP-hard problem. Traditionally, this computation has been addressed by means of the design of algorithms that find, by a clever exhaustive search, a linear combination of some generating matrix rows that provides a codeword with minimum weight. Therefore, as the dimension of the code or the size of the underlying finite field increase, so it does exponentially the run time. In this work, we prove that, given a generating matrix, there exists a column permutation which leads to a reduced row echelon form containing a row whose weight is the code distance. This result enables the use of permutations as representation scheme, in contrast to the usual discrete representation, which makes the search of the optimum polynomial time dependent from the base field. In particular, we have implemented genetic and CHC algorithms using this representation as a proof of concept. Experimental results have been carried out employing codes over fields with two and eight elements, which suggests that evolutionary algorithms with our proposed permutation encoding are competitive with regard to existing methods in the literature. As a by-product, we have found and amended some inaccuracies in the Magma Computational Algebra System concerning the stored distances of some linear codes.

Keywords: linear codes, minimum distance, genetic algorithms

1. INTRODUCTION

This talk is based in some results published in Cuéllar et al. (2021).

The computation of a word with minimum weight for a linear code is not an easy task. Vardy showed in Vardy (1997) that the decision problem associated to the computation of the minimum distance of a binary linear code is NP-complete. Hence, unless P = NP, we cannot expect to find a general polynomial time exact algorithm to compute the distance of an arbitrary linear code. This feature has been used to develop Code-based Cryptography, see D. J. Bernstein (2009), as one of main research lines looking for a possible new standard concerning Post-quantum Cryptography.

We can find different approaches in the literature to overcome the distance calculation. One is to design the linear code $\mathcal{C}$ subject to certain constraints using higher algebraic structures, to guarantee a lower bound for the distance $d(\mathcal{C})$. These approaches do not tackle the problem of finding the true distance, which remains unknown, although they ensure an error correction capability given by its precalculated lower bound.

Another kind of approaches focuses on finding the true distance using exact algorithms. Here, the problem is considered as a search procedure over a solution space, using heuristics to guide the search and to reduce the computational time of the algorithms. The most known algorithm was designed by Brouwer and Zimmermann, see Betten et al. (2006), and later extended by Lisonek and Trummer (2016), and it essays to build information sets from a generating matrix. The main drawback is that the efficiency of these methods is still non-polynomial, since the whole solution space has to be explored in the worst case. They have been successfully applied over small binary codes, but their computational time increases exponentially with the length of the code and the bit-size of the underlying finite field.

A third class of attemtps focuses on providing approximate methods to find lower and/or upper bounds of the distance. In this category, one of the first algorithms was provided in Leon (1988). Metaheuristics have also been used to solve the problem. Here, the methodology consists on finding an optimal solution $m^* \in \mathcal{A}^k$ that minimizes the fitness $f(m^*) = min_m\{w(mG)\}$ subject to the restriction $f(m^*) > 0$ (or equivalently, $m^* \neq 0$), where w denotes the Hamming weight. These methods provide an upper bound of $d(\mathcal{C})$. Some proposals we have found in the literature to tackle the problem were Genetic Algorithms, see Dontas

⋆ Research partially supported by SRA (State Research Agency / 10.13039/501100011033) under Grant No. PID2019-110525GB-I00.

and De Jong (1990), Simulated Annealing, see Muxiang and Fulong (1994), Ant Colony Optimization, see Bland (2007) and Bouzkraoui et al. (2018), for example.

One of the main limitations in many of the approximate methods is the high selective pressure over the non-valid trivial message $s = (0, 0, \ldots, 0)$, which leads to the fitness $f(s) = 0$. This problem has been highlighted and studied by some authors, as for instance in Berkani et al. (2015), and more recently in Berkani et al. (2017).

In this work we propose to address this problem from a completely new perspective. We prove that, given a generating matrix, there exists a column permutation which leads to a reduced row echelon form containing a row whose weight is the code distance. This result allows us to provide a novel problem statement and, therefore, new search mechanisms could be applied to provide a problem solution. In our case, we propose to use a permutation representation of the solution space to search the minimum distance. Hence a Generational Genetic Algorithm (GGA) and a Cross generational elitist selection, Heterogeneous recombination, and Cataclysmic mutation algorithm (CHC) are designed for solving the problem.

This has allowed us to discover some inaccuracies in the Magma Computational Algebra System concerning the stored distances of some linear codes.

2. PERMUTATION-BASED SCHEME

Let $0 < k \leq n$ be two non negative integers, and $\mathbb{F}_q$ the field with q elements. Our new perspective to compute the minimum distance relies in the the following result:

Theorem 1. Let G be a $k \times n$ generating matrix of a $[n, k]_q$-linear code $\mathcal{C}$ over the finite field $\mathbb{F}_q$. There exists a permutation $\sigma \in \mathcal{S}_n$ such that the RREF, R, of GP_σ, where P_σ is the permutation matrix of σ, satisfies that the Hamming weight of some of its rows reaches the minimum distance of $\mathcal{C}$. Consequently, if b is a row of R verifying such property, then bP_σ^{-1} is a non zero codeword of $\mathcal{C}$ with minimal weight.

By Theorem, finding the minimum distance of an $[n, k]_q$-linear code is reduced to find the minimum of the map $\mathfrak{d} : \mathcal{S}_n \to \mathbb{N}$ defined by

$$\mathfrak{d}(\sigma) = \min\{\mathrm{w}(b) \mid b \text{ is a row of the RREF of } GP_\sigma\} \quad (1)$$

for any $\sigma \in \mathcal{S}_n$. This encoding is then invariant with respect to the base field. Obviously, the computation of $\mathfrak{d}(\sigma)$, for some permutation σ, does depend on q and n.

3. ALGORITHMS

The Generational Genetic Algorithm (GGA) follow the classic genetic algorithm with elitism and reinitialization Back et al. (1997).

By Theorem 1, we are looking for permutations in order to compute the weights of the rows in the RREF of the permuted generating matrix. The algorithm works as follows: A population $P(t) \subseteq \mathcal{S}_n$ is initialized and evaluated with N random solutions at iteration $t = 0$. Then, the main loop of the algorithm is executed until a stopping condition is met. In this paper, the stopping criterion is to evaluate a maximum number of solutions so that the algorithms can be compared in performance. In order to test the algorithms, an additional stopping criterion has been included when a solution in the population reaches a previously known lower bound of the distance.

The loop of the algorithms start by selecting N parents according to the binary tournament selection operator Blickle and Thiele (1997). A crossover operator is applied to two parents to generate a pair of new solutions with probability p_c. If they are not combined, the mutation operator acts on the parents to generate two mutated solutions. All the N new solutions generated by either crossover or mutation form the population at the next iteration $P(t + 1)$. Finally, the solutions in $P(t + 1)$ are evaluated. An elitism component is included before the next iteration starts: If no solution in $P(t+1)$ has a fitness equal or better than the best in $P(t)$, then then worst in $P(t + 1)$ is replaced with the best in $P(t)$. Also, we include a reinitialization of $P(t + 1)$ with N new random solutions after $MaxReinit$ solution evaluations with no improvement in the fitness of the best solution found.

The CHC algorithm is an evolutionary algorithm whose initial version was proposed for binary encoding Eshelman (1991). This algorithm holds a balance between genotypic diversity in the solutions of the population, and convergence to local optima. It is based on four main components: elitist selection, the HUX solution recombination operator, an incest prevention check to avoid the recombination of similar solutions, and a population reinitialization method when a local optimum is found. Later versions of this algorithm are proposed for real and permutation encoding in Eshelman and Schaffer (1993); Cordón et al. (2006); Simões and Costa (2011). The adaptation of this algorithm it is mainly inspired in the proposals of Cordón et al. (2006); Simões and Costa (2011).

The distance between two solutions x, y in the population is computed using the Hamming distance, and the decrement dec of the crossover is updated as a percentage τ of the maximum Hamming distance between individuals in the population, where $\tau \in [0, 1]$ is an update rate, an input parameter to the algorithm.

The algorithm starts by initializing a population $P(t) \subseteq \mathcal{S}_n$ with N random solutions. Then, the average and maximum distances between all solutions are computed. The crossover threshold d is assigned to the average distance, and a threshold update rate dec is initialized to τ multiplied by the maximum distance. The main loop of the algorithm finishes when the aforementioned stopping condition is met. It works as follows: Firstly, the solutions in $P(t)$ are randomly shuffled and matched by pairs. These pairs of solutions are the parents to be combined. Then, the crossover operator is applied to each pair of parents to generate two offsprings, only if the distance between the two parents is not under the distance threshold d. If so, the offsprings are evaluated, and the population at the next iteration $P(t + 1)$ is created containing the best N solutions coming from $P(t)$ and the new generated solutions by crossover. If $P(t+1)$ is the same as $P(t)$, the crossover threshold d is decreased by dec. Only when d is zero or under zero, the population is reinitialized. In our

work, the reinitialization procedure replaces $P(t+1)$ with $N-1$ randomly generated solutions, and the best solution in $P(t)$. The values d and dec are recalculated for this new population.

4. EXPERIMENTATION

The parameters used in the experimentation are shown in Table 1. The row *Mutation probability by gene* describes the probability of mutation of each coordinate in a solution. The row *Evaluations* stands for the main stopping criterion of each algorithm (i.e. to reach a maximum number of evaluations). The row *Reinitialization* shows the number of solutions evaluated with no improvements required to reinitialize the population in GGA proposals. A secondary stopping criterion was set for all algorithms: if the true known distance of the code being evaluated (or a lower bound) is found, the algorithm stops even if the main stopping criterion is not satisfied yet.

Parameter	GGA-Order	CHC-Order
Population size	400	400
Crossover	AX_2	AX_2
Mutation	2-swap	–
Crossover probability	0.8	–
Mutation probability by gene	–	–
Evaluations	500000	500000
Reinitialization	100000	–

Table 1. Algorithms' parameters for experimentation with codes over $\mathbb{F}_8$

Tables 2 and 3 show the results obtained for each dataset containing a code over $\mathbb{F}_8$. The columns **Dataset** display the parameters of the corresponding best known linear codes according to Grassl (2007). Columns **B** describe the minimum weight found by each algorithm, and the subindices indicate the number of experiments that provided such result. Columns **W** introduce the worst minimum weight obtained. Columns **M** show the average of the minimum weights obtained in the 100 experiments. Finally, Columns **T/E** print the average time in seconds of each experiment, together with the average number of evaluations required to finish it.

5. ANALYSIS OF THE CODE $(75, 45, 17-27)$ OVER $\mathbb{F}_8$

According to Grassl (2007) and Bosma et al. (1997), the best known linear code over $\mathbb{F}_8$ with length 75 and dimension 45 is $(75, 45, 17-27)$ as it can be verified executing the first 7 lines of the MAGMA source code in Figure 1. However, our algorithms obtain the column permutation in (2), which provides a matrix whose RREF contains a row whose weight is 15. A codeword with this weight is exhibited in (3). MAGMA source code to verify our findings is set in lines 9 to 13 in Figure 1, which can be tested in the online version of MAGMA at the URL `http://magma.maths.usyd.edu.au/calc/`. Thus, the minimum distance is less or equal to 15, which is below of the lower bound provided by MAGMA.

$$(53, 11, 58, 36, 17, 27, 44, 8, 73, 24, 20, 71, 69, 46, 10, 43, 26, 61, 29, 57, 23, 6, 5, 67, 14, 4, 50, 45, 72, 59, 18, 25, 47, 28, 51, 68, 22, 48, 52, 42, 60, 21, 38, 64, 16, 34, 2, 3, 31, 33, 49, 63, 74, 54, 62, 19, 7, 1, 0, 15, 13, 66, 39, 65, 41, 30, 12, 37, 32, 70, 56, 40, 55, 9, 35) \quad (2)$$

$$(a^2+a, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, a, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, a, 0, 0, 0, 0, 0, 0, a^2+a, a^2, 0, a^2+a+1, 1, 0, a+1, 0, 0, a, 0, 0, 0, 0, 0, a^2+a+1, 0, 0, 0, a+1, 1, 0, 0, 0, 0, 0, a^2+a+1, 0, 0, 0, 0, 0, 0, a+1, 0, 0, 0, 0) \quad (3)$$

According to Grassl (2007), the linear code $(75, 45, 17-27)$ is constructed from the code with parameters $(77, 46, 18)$ described in Xing and Ling (2000). By additional experi-

	GGA (order)			
Dataset	**B**	**W**	**M**	**T / E**
$(30, 16, 10-13)$	$10_{(100)}$	10	$10_{(1)}$	$0.003_{(1)}/1.6$
$(30, 18, 9-11)$	$9_{(100)}$	9	$9_{(1)}$	$0.004_{(1)}/2$
$(45, 22, 15-21)$	$15_{(100)}$	15	$15_{(1)}$	$0.009_{(1)}/57.8$
$(45, 24, 14-19)$	$14_{(100)}$	14	$14_{(1)}$	$0.01_{(1)}/13.4$
$(45, 26, 12-17)$	$12_{(100)}$	12	$12_{(1)}$	$0.01_{(1)}/9.5$
$(45, 28, 11-15)$	$11_{(100)}$	11	$11_{(1)}$	$0.01_{(1)}/1.9$
$(60, 28, 21-28)$	$21_{(100)}$	21	$21_{(1)}$	$0.017_{(1)}/42.8$
$(60, 30, 20-27)$	$20_{(100)}$	20	$20_{(1)}$	$0.018_{(1)}/5$
$(60, 32, 19-25)$	$19_{(100)}$	19	$19_{(1)}$	$0.02_{(1)}/8.3$
$(60, 34, 17-23)$	$17_{(100)}$	17	$17_{(1)}$	$0.021_{(1)}/18.4$
$(75, 30, 28-40)$	$28_{(100)}$	28	$28_{(1)}$	$0.028_{(1)}/19.1$
$(75, 35, 24-35)$	$24_{(100)}$	24	$24_{(1)}$	$0.134_{(1)}/1716.2$
$(75, 40, 20-31)$	$20_{(100)}$	20	$20_{(1)}$	$0.036_{(1)}/5.9$
$\mathbf{(75, 45, 17-27)}$	$\mathbf{15}_{(100)}$	**15**	$\mathbf{15}_{(1)}$	$12.713_{(2)}/77303.3$
$(90, 19, 49-63)$	$49_{(100)}$	49	$49_{(1)}$	$0.252_{(2)}/5249.6$
$(90, 50, 21-35)$	$22_{(28)}$	24	$22.8_{(1)}$	$38.437_{(3)}/239954.9$
$(90, 60, 16-26)$	$16_{(99)}$	17	$16_{(1)}$	$15.42_{(3)}/86775.2$
$(90, 70, 11-17)$	$11_{(100)}$	11	$11_{(1)}$	$0.066_{(2)}/48.9$
$(130, 85, 23-40)$	$23_{(2)}$	26	$25_{(1)}$	$145.695_{(4)}/293522$
$(130, 95, 18-30)$	$18_{(80)}$	19	$18.2_{(1)}$	$87.796_{(3)}/177852.8$

Table 2. Results of GGA-Order over $\mathbb{F}_8$ codes

	CHC (order)			
Dataset	**B**	**W**	**M**	**T / E**
$(30, 16, 10-13)$	$10_{(100)}$	10	$10_{(1)}$	$0.004_{(2)}/1.6$
$(30, 18, 9-11)$	$9_{(100)}$	9	$9_{(1)}$	$0.005_{(2)}/2$
$(45, 22, 15-21)$	$15_{(100)}$	15	$15_{(1)}$	$0.01_{(2)}/57.8$
$(45, 24, 14-19)$	$14_{(100)}$	14	$14_{(1)}$	$0.011_{(2)}/13.4$
$(45, 26, 12-17)$	$12_{(100)}$	12	$12_{(1)}$	$0.011_{(2)}/9.5$
$(45, 28, 11-15)$	$11_{(100)}$	11	$11_{(1)}$	$0.011_{(2)}/1.9$
$(60, 28, 21-28)$	$21_{(100)}$	21	$21_{(1)}$	$0.019_{(2)}/42.8$
$(60, 30, 20-27)$	$20_{(100)}$	20	$20_{(1)}$	$0.019_{(2)}/5$
$(60, 32, 19-25)$	$19_{(100)}$	19	$19_{(1)}$	$0.021_{(2)}/8.3$
$(60, 34, 17-23)$	$17_{(100)}$	17	$17_{(1)}$	$0.021_{(2)}/18.4$
$(75, 30, 28-40)$	$28_{(100)}$	28	$28_{(1)}$	$0.029_{(2)}/19.1$
$(75, 35, 24-35)$	$24_{(100)}$	24	$24_{(1)}$	$0.12_{(1)}/1269.2$
$(75, 40, 20-31)$	$20_{(100)}$	20	$20_{(1)}$	$0.037_{(2)}/5.9$
$\mathbf{(75, 45, 17-27)}$	$\mathbf{15}_{(100)}$	**15**	$\mathbf{15}_{(1)}$	$11.695_{(2)}/72438.6$
$(90, 19, 49-63)$	$49_{(100)}$	49	$49_{(1)}$	$0.179_{(1)}/3279.7$
$(90, 50, 21-35)$	$22_{(34)}$	24	$22.7_{(1)}$	$41.898_{(4)}/267888.3$
$(90, 60, 16-26)$	$16_{(100)}$	16	$16_{(1)}$	$13.265_{(3)}/78140$
$(90, 70, 11-17)$	$11_{(100)}$	11	$11_{(1)}$	$0.066_{(1)}/48.9$
$(130, 85, 23-40)$	$23_{(1)}$	26	$24.9_{(1)}$	$63.325_{(3)}/135700$
$(130, 95, 18-30)$	$18_{(88)}$	19	$18.1_{(1)}$	$84.811_{(3)}/185147.7$

Table 3. Results of CHC-Order over $\mathbb{F}_8$ codes

Fig. 1. MAGMA source code exhibiting a codeword with weight 15 for the code $(75, 45, 17 - 27)$

```
/* Build the code */
F<a> := GF(8);
V := VectorSpace(F,75);
C := BKLC(F,75,45);
C:Minimal;
/* Show the lower and upper bounds for the code */
BKLCLowerBound(F,75,45), BKLCUpperBound(F,75,45);
/* Set the word of weight 15 */
c := V ! [a^2 + a, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, a, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, a, 0, 0, 0, 0, 0, 0, a^2 + a, a^2, 0,
a^2 + a + 1, 1, 0, a + 1, 0, 0, a, 0, 0, 0, 0, 0,
a^2 + a + 1, 0, 0, 0, a + 1, 1, 0, 0, 0, 0, 0,
a^2 + a + 1, 0, 0, 0, 0, 0, 0, a + 1, 0, 0, 0, 0];
/* Checks if the word c belongs to code C */
c in C;
/* Prints the weight of the codeword c in code C */
Weight(c);
```

ments we have found a codeword of weight 15 in this code. Additional tests were performed for other codes considered in Xing and Ling (2000). More specifically, we calculated permutations for the codes $(75, 44, 18)$ and $(76, 45, 18)$, respectively, which helps us finding the corresponding codewords both with weight 15.

REFERENCES

Back, T., Fogel, D.B., and Michalewicz, Z. (1997). *Handbook of Evolutionary Computation*. IOP Publishing Ltd., GBR, 1st edition.

Berkani, A., Azouaoui, A., and Belkasmi, M. (2015). Soft-decision decoding by a compact genetic algorithm using higher selection pressure. In *2015 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, 1–6.

Berkani, A., Azouaoui, A., Belkasmi, M., Aylaj, B., University, M.V.R., and labo, S. (2017). Improved decoding of linear block codes using compact genetic algorithms with larger tournament size. *International Journal of Computer Science Issues*, 14(1).

Betten, A., Braun, M., Fripertinger, H., Kerber, A., Kohnert, A., and Wassermann, A. (2006). *Error-Correcting Linear Codes*, volume 18 of *Algorithms and Computation in Mathematics*. Springer.

Bland, J. (2007). Local search optimisation applied to the minimum distance problem. *Advanced Engineering Informatics*, 21(4), 391 – 397. doi: https://doi.org/10.1016/j.aei.2007.01.002.

Blickle, T. and Thiele, L. (1997). A comparison of selection schemes used in evolutionary algorithms. *Evolutionary Computation*, 4, 361–394.

Bosma, W., Cannon, J., and Playoust, C. (1997). The Magma algebra system. I. The user language. *J. Symbolic Comput.*, 24(3-4), 235–265. doi:10.1006/jsco.1996.0125. URL `http://dx.doi.org/10.1006/jsco.1996.0125`. Computational algebra and number theory (London, 1993).

Bouzkraoui, H., Azouaoui, A., and Hadi, Y. (2018). New ant colony optimization for searching the minimum distance for linear codes. In *2018 International Conference on Advanced Communication Technologies and Networking (CommNet)*, 1–6.

Cordón, O., Damas, S., and Santamaría, J. (2006). Feature-based image registration by means of the chc evolutionary algorithm. *Image and Vision Computing*, 24(5), 525 – 533.

Cuéllar, M.P., Gómez-Torrecillas, J., Lobillo, F.J., and Navarro, G. (2021). Genetic algorithms with permutation-based representation for computing the distance of linear codes. *Swarm and Evolutionary Computation*, 60, 100797. doi: https://doi.org/10.1016/j.swevo.2020.100797.

D. J. Bernstein, J. Buchmann, E.D. (2009). *Post-Quantum Cryptography*. Springer-Verlag Berlin Heidelberg.

Dontas, K. and De Jong, K. (1990). Discovery of maximal distance codes using genetic algorithms. In *[1990] Proceedings of the 2nd International IEEE Conference on Tools for Artificial Intelligence*, 805–811. doi: 10.1109/TAI.1990.130442.

Eshelman, L.J. (1991). The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. In G.J. RAWLINS (ed.), *Foundations of Genetic Algorithms*, volume 1 of *Foundations of Genetic Algorithms*, 265 – 283. Elsevier. doi: https://doi.org/10.1016/B978-0-08-050684-5.50020-3.

Eshelman, L.J. and Schaffer, J.D. (1993). Real-coded genetic algorithms and interval-schemata. In *Foundations of Genetic Algorithms*, volume 2 of *Foundations of Genetic Algorithms*, 187 – 202. Elsevier.

Grassl, M. (2007). Bounds on the minimum distance of linear codes and quantum codes. Online available at `http://www.codetables.de`. Accessed 2020-01-22.

Leon, J.S. (1988). A probabilistic algorithm for computing minimum weights of large error-correcting codes. *IEEE Transactions on Information Theory*, 34(5), 1354–1359. doi:10.1109/18.21270.

Lisonek, P. and Trummer, L. (2016). Algorithms for the minimum weight of linear codes. *Adv. in Math. of Comm.*, 10, 195–207.

Muxiang, Z. and Fulong, M. (1994). Simulated annealing approach to the minimum distance of error-correcting codes. *International Journal of Electronics - INT J ELECTRON*, 76, 377–384. doi: 10.1080/00207219408925934.

Simões, A. and Costa, E. (2011). Memory-based CHC algorithms for the dynamic traveling salesman problem. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, GECCO '11, 1037–1044. Association for Computing Machinery, New York, NY, USA. doi:10.1145/2001576.2001717. URL `https://doi.org/10.1145/2001576.2001717`.

Vardy, A. (1997). The intractability of computing the minimum distance of a code. *IEEE Transactions on Information Theory*, 43(6), 1757–1766. doi: 10.1109/18.641542.

Xing, C. and Ling, S. (2000). A class of linear codes with good parameters from algebraic curves. *IEEE Transactions on Information Theory*, 46(4), 1527–1532.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Event-triggered observer design of nonlinear systems with multiple sensor nodes. Extended abstract ⋆

E. Petri *R. Postoyan *D. Astolfi **D. Nešić ***W.P.M.H. Heemels ****

** Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France. (e-mail: elena.petri@univ-lorraine.fr, romain.postoyan@univ-lorraine.fr)*
*** Université Claude Bernard Lyon 1, CNRS, LAGEPP UMR 5007, F-69100, Villeurbanne, France. (e-mail: daniele.astolfi@univ-lyon1.fr)*
**** Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, 3010 Victoria, Australia. (e-mail: dnesic@unimelb.edu.au)*
***** Department of Mechanical Engineering, Eindhoven University of Technology, The Netherlands. (e-mail: m.heemels@tue.nl)*

Abstract: We investigate the scenario where a perturbed nonlinear system communicates its output measurements to a remote observer via a network. The sensors are grouped into N nodes and each of these nodes decides when its measured data is transmitted over the network independently. Given a (continuous-time) observer, we present an approach to design local (dynamic) transmission policies to obtain accurate state estimates, while only sporadically using the communication network. We prove a practical convergence property to the origin for the estimation error and we show there exists a uniform strictly positive minimum inter-event time for each local triggering rule under mild conditions on the plant. The analysis relies on hybrid Lyapunov tools. The efficiency of the proposed techniques is illustrated on a numerical case study of a flexible robotic arm.

Preamble: *This extended abstract is based on the journal submission Petri et al. (2022), which is in agreement with the conference policy.*

While digital networks exhibit a range of benefits for control applications in terms of ease of installation, maintenance and reduced weight and volume, they also require adapted control theoretical tools to cope with the induced communication constraints (e.g., sampling, delays, packet drops, scheduling, quantization), see e.g., Hespanha et al. (2007); Heemels and Van De Wouw (2010). In this work, we concentrate on the state estimation of nonlinear systems over a digital channel and we focus on the effect of sampling. In particular, we consider state estimation where the plant is nonlinear, perturbed and communicates its measurements over a digital network to a remote observer, whose goal is to estimate the plant state, see Fig. 1.

The communication schedule is very important to guarantee good estimation performance. An option is to generate the transmission instants based on time, in which case we talk of time-triggered strategies for which various results are available in the literature, see, e.g., Postoyan and Nešić (2011); Li et al. (2017); Ferrante et al. (2016); Mazenc et al. (2015); Dačić and Nešić (2008). However, this paradigm may generate (significantly) more transmissions over the network than necessary to fulfil the estimation task, thereby leading to a waste of the network resources. As a potential and promising solution, one can use event-triggered transmissions to overcome this drawback, see e.g., Heemels et al. (2012) and references therein. In this case, an event-based triggering rule monitors the plant measurement and/or the observer state and decides when an output transmission is needed.

Various event-triggered techniques are available in the literature for estimation, see, e.g., Scheres et al. (2021); Li et al. (2010); Shi et al. (2014); Li and Lemmon (2011); Trimpe (2014); Yu et al. (2021); Song et al. (2021); Shi et al. (2016); Huang et al. (2019); Sijs and Lazar (2012); Hu et al. (2020); Etienne and Di Gennaro (2016); Etienne et al. (2017a,b); Tong et al. (2020); Niu et al. (2020). A dominant approach consists in implementing a copy of the observer within the sensor and then use its information to define the transmission instants, see e.g., Scheres et al. (2021); Li et al. (2010); Shi et al. (2014); Li and Lemmon (2011); Trimpe (2014); Yu et al. (2021); Song et al. (2021). A possible drawback then is that it may require significant computation capabilities on the sensors, especially in the case of large-scale systems, or highly nonlinear dynamics, which may be unavailable. Another possible solution is to follow an event-triggered strategy, which is only based on a static condition involving the measured output and its

⋆ This work was funded by Lorraine Université d'Excellence LUE, HANDY project ANR-18-CE40-0010-02, the France Australian collaboration project IRP-ARS CNRS and the Australian Research Council under the Discovery Project DP200101303.

past transmitted value(s) see, e.g., Shi et al. (2016); Huang et al. (2019); Sijs and Lazar (2012); Hu et al. (2020); Etienne et al. (2017a); Etienne and Di Gennaro (2016); Etienne et al. (2017b); Tong et al. (2020). Consequently, it is not necessary to implement a copy of the observer in the sensors and thus the sensors are not required to have significant computation capabilities. However, such static triggering rules may generate a lot of transmissions, moreover the results in Shi et al. (2016); Huang et al. (2019); Sijs and Lazar (2012); Hu et al. (2020); Etienne and Di Gennaro (2016); Etienne et al. (2017a,b); Tong et al. (2020) only apply to specific classes of systems and a centralized scenario, where all sensors communicate simultaneously over the network, with the exception of Shi et al. (2016) and Hu et al. (2020).

In this work, we adopt an event-triggered approach based only on the measured output and the last transmitted output value. This strategy keeps monitoring the plant output, and thereby may lead to less transmissions compared to a self-triggering approach, and it does not require a copy of the observer, which simplifies the implementation and requires less computation capability on the sensor. The main novelties are, first, the design of a new triggering rule, which involves an auxiliary scalar variable for each sensor node, that will have several benefits. Second, the proposed results apply to general, perturbed nonlinear systems contrary to the vast majority of works in the literature, which concentrates on specific classes of systems, see e.g., Li et al. (2010); Shi et al. (2014); Li and Lemmon (2011); Trimpe (2014); Yu et al. (2021); Song et al. (2021); Shi et al. (2016); Huang et al. (2019); Sijs and Lazar (2012); Hu et al. (2020); Etienne and Di Gennaro (2016); Etienne et al. (2017a,b); Tong et al. (2020); Niu et al. (2020). Third, the triggering strategies are decentralized. Indeed, we consider the scenario with N sensor nodes, where each node decides independently when to transmit its local data to the observer via a digital network. Consequently, each sensor node has its own triggering rule and several nodes are allowed to communicate at the same time instant. The considered setup is depicted in Fig. 1.

Our design is following an emulation-based approach in the sense that the observer is first designed ignoring the effects of the communication network. In particular, we assume that the observer has been designed in continuous-time in such a way that it satisfies an input-to-state stability property, that holds for many observer design techniques of the literature, see e.g., Astolfi et al. (2021); Shim and Liberzon (2015) and the references therein. Afterwards, we take the network into account and consequently the observer knows only the networked version of the output, which is generated using a zero-order-hold device between two successive transmission instants. We then design a triggering rule for each sensor node to approximately preserve the original properties of the observer. As already stated, the triggering rules are dynamic in the sense that they involve a local scalar auxiliary variable, which essentially filters an absolute threshold type condition, see e.g., Etienne et al. (2017a,b); Etienne and Di Gennaro (2016); Tong et al. (2020). This is a new in the context of estimation, to the best of the authors knowledge, and is inspired by related event-triggering control techniques, see e.g., Girard (2014); Tanwani et al. (2015); Tabuada (2007). Importantly, there is no need to implement a copy of the observer at each sensor node, which has clear computational advantages. Indeed, each sensor just needs to know the difference between its current output and its last transmitted output value and its local auxiliary scalar variable. We model the overall system as an hybrid system using the formalism of Goebel et al. (2012); Heemels et al. (2021), where a jump corresponds to a transmission of the current output measured by one of the sensor nodes to the observer.

To design the event-triggered estimation scheme, we provide easily usable conditions on the triggering rules, which ensure that the estimation error system satisfies a global practical stability property. The analysis relies on hybrid Lyapunov tools, see Goebel et al. (2012). Note that, we do not guarantee an asymptotic stability property, but only a practical one in general, which is a consequence of the absence of a copy of the observer in the triggering mechanism. Afterwards, we prove that there exists a uniform strictly positive minimum time between any two successive transmissions for each local triggering rule under mild boundedness conditions on the plant state and its input, thereby excluding the Zeno phenomena. Moreover, we provide explicit conditions under which the proposed event-triggered observer stops transmitting, which it is a clear advantage against time-triggered strategies. The proposed results can then be extended in various ways (to be robust to additive measurement noise, to the case where the plant input is also triggered etc.). Finally, we will illustrate the efficiency of the approach in a numerical case study of a flexible robotic arm.

This work is an extension of the preliminary conference paper Petri et al. (2021), where only linear time-invariant systems and a centralized transmission strategy were considered.

REFERENCES

Astolfi, D., Alessandri, A., and Zaccarian, L. (2021). Stubborn and dead-zone redesign for nonlinear observers and filters. *IEEE Transactions on Automatic Control*, 66(2), 667–682.

Dačić, D. and Nešić, D. (2008). Observer design for wired linear networked control systems using matrix inequalities. *Automatica*, 44(11), 2840–2848.

Etienne, L. and Di Gennaro, S. (2016). Event-triggered observation of nonlinear Lipschitz systems via impulsive observers. *IFAC-PapersOnLine*, 49(18), 666–671.

Etienne, L., Di Gennaro, S., and Barbot, J.P. (2017a). Periodic event-triggered observation and control for nonlinear Lipschitz systems using impulsive observers. *International Journal of Robust and Nonlinear Control*, 27(18), 4363–4380.

Etienne, L., Khaled, Y., Di Gennaro, S., and Barbot, J.P. (2017b). Asynchronous event-triggered observation and control of linear systems via impulsive observers. *Journal of the Franklin Institute*, 354(1), 372–391.

Ferrante, F., Gouaisbaut, F., Sanfelice, R.G., and Tarbouriech, S. (2016). State estimation of linear systems in the presence of sporadic measurements. *Automatica*, 73, 101–109.

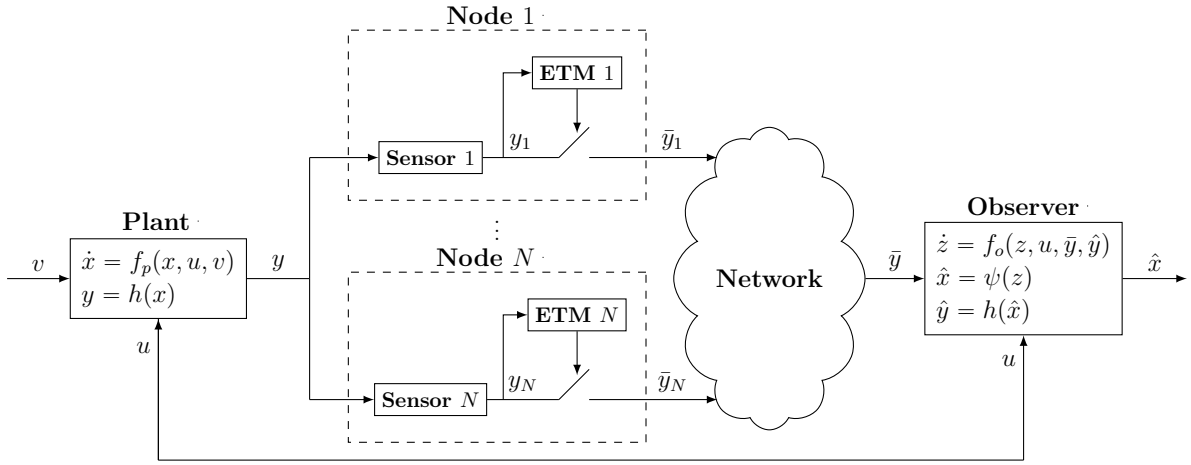


Fig. 1. Block diagram representing the system architecture (ETM: Event-Triggering Mechanism) Petri et al. (2022), where $x \in \mathbb{R}^{n_x}$ is the state to be estimated, $u \in \mathbb{R}^{n_u}$ is the measured input, $y := (y_1, \ldots, y_N) \in \mathbb{R}^{n_y} := \mathbb{R}^{n_{y1}} \times \cdots \times \mathbb{R}^{n_{yN}}$ is the measured output, where y_i, with $i \in \{1, \ldots, N\}$ is the output measured by sensor i, $v \in \mathbb{R}^{n_v}$ is an unmeasured disturbance input, $z \in \mathbb{R}^{n_z}$ is the observer state with $n_z \geqslant n_x$, $\hat{x} \in \mathbb{R}^{n_x}$ is the state estimate, $\bar{y} := (\bar{y}_1 \ldots \bar{y}_N) \in \mathbb{R}^{n_y} := \mathbb{R}^{n_{y1}} \times \cdots \times \mathbb{R}^{n_{yN}}$ is the networked version of the output y.

Girard, A. (2014). Dynamic triggering mechanisms for event-triggered control. *IEEE Transactions on Automatic Control*, 60(7), 1992–1997.

Goebel, R., Sanfelice, R.G., and Teel, A.R. (2012). *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*. Princeton University Press, New Jersey, U.S.A.

Heemels, W.P.M.H., Bernard, P., Scheres, K.J.A., Postoyan, R., and Sanfelice, R.G. (2021). Hybrid systems with continuous-time inputs: Subtleties in solution concepts and existence properties. *IEEE Conference on Decision and Control,* Austin, USA, 5361–5366.

Heemels, W.P.M.H., Johansson, K.H., and Tabuada, P. (2012). An introduction to event-triggered and self-triggered control. *IEEE Conference on Decision and Control,* Maui, HI, USA, 3270–3285.

Heemels, W.P.M.H. and Van De Wouw, N. (2010). Stability and stabilization of networked control systems. In *Networked Control Systems*, 203–253. Springer.

Hespanha, J.P., Naghshtabrizi, P., and Xu, Y. (2007). A survey of recent results in networked control systems. *Proceedings of the IEEE*, 95(1), 138–162.

Hu, J., Wang, Z., Liu, G.P., Jia, C., and Williams, J. (2020). Event-triggered recursive state estimation for dynamical networks under randomly switching topologies and multiple missing measurements. *Automatica*, 115.

Huang, J., Shi, D., and Chen, T. (2019). Robust event-triggered state estimation: A risk-sensitive approach. *Automatica*, 99, 253–265.

Li, L. and Lemmon, M. (2011). Performance and average sampling period of sub-optimal triggering event in event triggered state estimation. *IEEE Conference on Decision and Control and European Control Conference,* Orlando, USA, 1656–1661.

Li, L., Lemmon, M., and Wang, X. (2010). Event-triggered state estimation in vector linear processes. *American Control Conference,* Baltimore, USA, 2138–2143.

Li, Y., Phillips, S., and Sanfelice, R.G. (2017). Robust distributed estimation for linear systems under intermittent information. *IEEE Transactions on Automatic Control*, 63(4), 973–988.

Mazenc, F., Andrieu, V., and Malisoff, M. (2015). Design of continuous–discrete observers for time-varying nonlinear systems. *Automatica*, 57, 135–144.

Niu, Y., Sheng, L., Gao, M., and Zhou, D. (2020). Dynamic event-triggered state estimation for continuous-time polynomial nonlinear systems with external disturbances. *IEEE Transactions on Industrial Informatics*, 17(6), 3962–3970.

Petri, E., Postoyan, R., Astolfi, D., Nešić, D., and Heemels, W.P.M.H. (2021). Event-triggered observer design for linear systems. *IEEE Conference on Decision and Control,* Austin, USA, 546–551.

Petri, E., Postoyan, R., Astolfi, D., Nešić, D., and Heemels, W.P.M.H. (2022). Decentralized event-triggered estimation of nonlinear systems. Submitted to *Automatica*.

Postoyan, R. and Nešić, D. (2011). A framework for the observer design for networked control systems. *IEEE Transactions on Automatic Control*, 57(5), 1309–1314.

Scheres, K.J.A., Chong, M.S.T., Postoyan, R., and Heemels, W.P.M.H. (2021). Event-triggered state estimation with multiple noisy sensor nodes. *IEEE Conference on Decision and Control,* Austin, TX, USA, 558–563.

Shi, D., Chen, T., and Darouach, M. (2016). Event-based state estimation of linear dynamic systems with unknown exogenous inputs. *Automatica*, 69, 275–288.

Shi, D., Chen, T., and Shi, L. (2014). Event-triggered maximum likelihood state estimation. *Automatica*, 50(1), 247–254.

Shim, H. and Liberzon, D. (2015). Nonlinear observers robust to measurement disturbances in an ISS sense. *IEEE Transactions on Automatic Control*, 61(1), 48–61.

Sijs, J. and Lazar, M. (2012). Event based state estimation with time synchronous updates. *IEEE Transactions on Automatic Control*, 57(10), 2650–2655.

Song, C., Wang, H., Tian, Y., and Zheng, G. (2021). Event-triggered observer design for output-sampled systems. *Nonlinear Analysis: Hybrid Systems*, 43, 101112.

Tabuada, P. (2007). Event-triggered real-time scheduling of stabilizing control tasks. *IEEE Transactions on Automatic Control*, 52(9), 1680–1685.

Tanwani, A., Teel, A., and Prieur, C. (2015). On using norm estimators for event-triggered control with dynamic output feedback. *IEEE Conference on Decision and Control,* Osaka, Japan, 5500–5505.

Tong, Y., Tong, D., Chen, Q., and Zhou, W. (2020). Finite-time state estimation for nonlinear systems based on event-triggered mechanism. *Circuits, Systems, and Signal Processing*, 1–21.

Trimpe, S. (2014). Stability analysis of distributed event-based state estimation. *IEEE Conference on Decision and Control,* Florence, Italy, 2013–2019.

Yu, H., Shang, J., and Chen, T. (2021). On stochastic and deterministic event-based state estimation. *Automatica*, 123, 109314.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Global synchronization of a tree-like network of Kuramoto oscillators (Extended abstract) ⋆

S. Mariano * **R. Bertollo** ** **R. Postoyan** * **L. Zaccarian** **,***

* *Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France (e-mail: simone.mariano@univ-lorraine.fr).*
** *Department of Industrial Engineering, University of Trento, Italy.*
*** *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France.*

Abstract: We study tree-like networks of leaderless Kuramoto-like non-identical oscillators having time-varying natural frequencies taking values in a compact set. We interconnect the oscillators via a novel class of hybrid coupling rules inducing uniform global practical asymptotic stability of the synchronization set, thereby ensuring global uniform convergence. Moreover, we show that the synchronization set is uniformly globally finite-time stable whenever the coupling function is discontinuous at the origin. Numerical simulation results illustrate the advantage of the proposed model with respect to non-uniform behavior typically found with classical Kuramoto models.

Keywords: Synchronization, hybrid dynamical systems, multi-agent systems, finite-time stability, Lyapunov methods.

This extended abstract presents the results given in the paper "Leaderless uniform global asymptotic and finite-time synchronization of Kuramoto-like oscillators" (Mariano et al. (2021)) submitted to Automatica for publication, as allowed by the conference guidelines.

Oscillatory behaviours and their studies have always been a topic of interest in engineering and science in general. To address this problem, the Kuramoto model (Kuramoto (1975)), while being proposed originally as a model to describe chemical and biological oscillators, has found widespread applications to describe a broad family of oscillatory behaviours Acebron et al. (2005). In Kuramoto (1975), the evolution of an "all-to-all" network, i.e., a fully connected graph, of n phase-coupled (possibly) heterogeneous oscillators is described as

$$\dot{\theta}_i = \omega_i + \frac{\kappa}{n} \sum_{j \in \mathcal{V}_i} \sin(\theta_j - \theta_i) \qquad i \in \{1, \dots, n\} \quad (1)$$

where $\theta_i \in \mathbb{R}$ is the phase of the i-th oscillator, $\mathcal{V}_i := \{1, \dots, n\} \setminus \{i\}$, $\omega_i \in \mathbb{R}$ is its natural frequency and $\kappa > 0$ is the gain of the coupling action between each pair of oscillators. Neuroscience (Tass (2003); Cumin and Unsworth (2007)), chemistry (Forrester (2015)) and electrical engineering (Dörfler and Bullo (2012)), to cite a few, are examples of contexts where the Kuramoto model has been successfully used. The control community has taken an active interest in the Kuramoto model and in the last two decades has provided rigorous guarantees of its synchronization properties in various settings, see, e.g., Jafarpour and Bullo (2019); Jadbabaie et al. (2004); Chopral and Spong (2009); Dörfler and Bullo (2011); Leonard et al. (2012); Aokii (2015); Sepulchre et al. (2007); Aeyels and Rogge (2004).

Among the phenomena characterizing oscillating systems, collective synchronization plays a key role. However, the results in the existing literature come with several shortcomings preventing this phenomenon from occurring. First, when the network comprises oscillators with the same natural frequency, i.e., when $\omega_i = \omega$ for all $i \in \{1, \dots, n\}$ in (1), it is now well-known that a system of Kuramoto oscillators admits, in addition to stable equilibria coinciding with the synchronization set, equilibria that are unstable (see, e.g., Strogatz (2000); Sepulchre et al. (2007)). The downside of this result is that the closer a solution is initialized to an unstable equilibrium, the longer it will take for phase synchronization to arise: we talk of *non-uniform* convergence Sepulchre et al. (2007). While non-uniform synchronization may naturally characterize certain physical (Oud (2006)) and biological systems, in general it is not a desirable property for engineering applications, for which we have the freedom to design the interconnection rules. Indeed, as a first drawback, the lack of uniformity may induce arbitrarily slow convergence to the attractor set and poor robustness properties (Miller and Pachter (1997)). Secondly, it may occur in the classical Kuramoto model that the angular phase mismatch between adjacent oscillators remains constant and different from zero indefinitely: in this case we talk of *phase locking* (Aeyels and Rogge (2004)), which hampers the capability to reach asymptotic collective synchronization. Thirdly, in critical applications, finite-time convergence, instead of only asymptotic synchronization, may be a mandatory requirement (Polyakov (2011)). To the authors' best knowledge, the case of finite-time synchronization of Kuramoto oscillators has been studied only in (Wu and

⋆ Work supported by the ANR under grant HANDY ANR-18-CE40-0010

Li (2018)), where the authors propose a multiplex control to synchronize non-identical oscillators in fixed-time. Finally, contrary to the case of non-identical oscillators with constant natural frequencies, which has been widely investigated in the literature (Dörfler and Bullo (2014)), oscillators with time-varying natural frequencies have been studied only in the settings of second order Kuramoto models, see, e.g., Dörfler and Bullo (2012), as far as we know.

In this context, we propose novel hybrid rules to interconnect oscillators, by exploiting the periodicity of phases. We consider an undirected tree graph $\mathcal{G}_u = (\mathcal{V}, \mathcal{E}_u)$ with $n = |\mathcal{V}|$ nodes and $m = |\mathcal{E}_u|$ edges and we assign an arbitrary orientation to $\mathcal{G}_u$, which leads to the oriented tree $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The continuous dynamics of each oscillator, represented by a node in $\mathcal{G}$, is defined as

$$\dot{\theta}_i = \omega_i(\mathfrak{t}) + \kappa \sum_{j \in \mathcal{O}_i} \sigma(\theta_j - \theta_i + 2q_{ij}\pi) - \kappa \sum_{j \in \mathcal{I}_i} \sigma(\theta_i - \theta_j + 2q_{ji}\pi), \qquad x \in C, \qquad (2)$$

where the sets $\mathcal{I}_i \cup \mathcal{O}_i := \mathcal{V}_i$ represent the *in-neighbors* and *out-neighbors* of node i. We consider each phase θ_i, $i \in \mathcal{V}$, evolving on $[-\pi - \delta, \pi + \delta]$, with $\delta \in (0, \pi)$. Variable q_{ij}, $(i, j) \in \mathcal{E}$, is a logic state, taking values in $\{-1, 0, 1\}$. The role of q_{ij} is to unwind the difference between the two phases θ_j and θ_i through jumps while it remains constant along flows. The time-varying parameter $\omega_i(\mathfrak{t}) \in [\omega_m, \omega_M]$ in (2) is the instantaneous frequency of oscillator i while the strictly positive gain $\kappa \in \mathbb{R}_{>0}$ scales the coupling between oscillators. The coupling action is expressed by σ, which is piecewise continuous on $\operatorname{dom}\sigma := [-\pi - \delta, \pi + \delta]$, and is selected such that $\sigma(s) = -\sigma(-s)$ for any $s \in \operatorname{dom}\sigma$ and that $\operatorname{sgn}(s)\sigma(s) \geq \alpha(|s|)$ for any $s \in \operatorname{dom}\sigma \setminus \{0\}$ and for some function α positive definite and non-decreasing. Examples of suitable functions σ are depicted in Figure 2, together with the sine function used in (1), for the sake of comparison (additional selections can be found in Mariano et al. (2021)). The state x collects all the q_{ij}, $(i, j) \in \mathcal{E}$, and θ_i, $i \in \mathcal{V}$. The flow set C in (2) is selected as the closed complement of the jump set D. Through jumps we ensure that the proposed model is well-defined: We design a first set of jump rules to guarantee that the argument $\theta_j - \theta_i + 2q_{ij}\pi$ of σ in (2) belongs to $\operatorname{dom}\sigma = [-\pi - \delta, \pi + \delta]$ when flowing. A second set of jump rules is introduced for when one of the oscillators $i \in \mathcal{V}$ reaches $|\theta_i| = \pi + \delta$. In this case, a jump of 2π is enforced so that the phase is mapped into $(-\pi - \delta, \pi + \delta)$ while remaining the same modulo 2π.

The novel class of hybrid coupling rules proposed to interconnect the oscillators induces uniform global practical asymptotic stability of the synchronization set, defined as

$$\mathcal{A} := \{x \in C \cup D : \theta_i = \theta_j + 2k_{ij}\pi, \forall (i, j) \in \mathcal{E}\}, \qquad (3)$$

thereby ensuring global uniform convergence, while still locally preserving the original behavior of Kuramoto oscillators for suitable selections of σ. The practical stability results can be summarized in the following theorems, whose proofs are given in Mariano et al. (2021).

Theorem 1. Given set $\mathcal{A}$ in (3), the following holds.

(i) All maximal solutions of the considered hybrid model are $\mathfrak{t}-$complete, i.e., their continuous time domain is unbounded;

(ii) there exists a class $\mathcal{KL}$ function $\beta_\circ$ and a class $\mathcal{K}$ gain $\gamma_\circ$, both of them independent of κ, such that, for any $\kappa > 0$, all solutions x satisfy

$$|x(\mathfrak{t}, \mathfrak{j})|_{\mathcal{A}} \leq \beta_\circ(|x(0, 0)|_{\mathcal{A}}, \kappa\mathfrak{t}) + \gamma_\circ((\kappa)^{-1}(|\omega_M - \omega_m|)), \quad (4)$$

for all $(\mathfrak{t}, \mathfrak{j}) \in \operatorname{dom} x$, where $\omega_M - \omega_m$ is the maximum mismatch of the instantaneous frequency of two adjacent oscillators. ■

In view of item (i) of Theorem 1, item (ii) of Theorem 1 provides an insightful bound (4) illustrating the trend of the continuous time evolution of the hybrid solutions to our model, and the role of κ in speeding up their transient and reducing their asymptotic disagreement. Hence, Theorem 1 implies that the oscillator phases uniformly converge to any desired neighborhood of $\mathcal{A}$ by taking κ sufficiently large, thus the practical nature of the result. A useful outcome of the mild regularity conditions that we require from σ is that defining σ to be discontinuous at the origin, as in the staircase function represented in Fig. 2, leads to a desirable sliding-like behavior of the solutions in the attractor $\mathcal{A}$. This sliding property induces interesting advantages of the behavior of solutions. A first advantage is that, even with non-uniform natural frequencies, we prove uniform global $\mathcal{KL}$ asymptotic stability of $\mathcal{A}$ for a large enough coupling gain κ, due to the well-known ability of sliding-mode mechanisms of dominating unknown additive bounded disturbances acting on the dynamics. A second advantage is that we can guarantee a finite-time convergence. Finally, by taking κ sufficiently large, we actually prove prescribed finite-time convergence (see Song et al. (2017)). These properties are formalized in the theorem below.

Theorem 2. If σ is discontinuous at the origin, then set $\mathcal{A}$ in (3) is prescribed finite-flowing-time stable, i.e., for each $T > 0$ there exists $\kappa^\star > 0$ such that for each $\kappa \geq \kappa^\star$:
(i) there exists $\beta \in \mathcal{KL}$ such that all solutions x satisfy $|x(\mathfrak{t}, \mathfrak{j})|_{\mathcal{A}} \leq \beta(|x(0, 0)|_{\mathcal{A}}, \mathfrak{t} + \mathfrak{j}), \quad \forall (\mathfrak{t}, \mathfrak{j}) \in \operatorname{dom} x$;

(ii) all solutions x satisfy, $x(\mathfrak{t}, \mathfrak{j}) \in \mathcal{A}$ for all $(\mathfrak{t}, \mathfrak{j}) \in \operatorname{dom} x$ with $\mathfrak{t} \geq T$. ■

Simulations are provided to illustrate the theoretical guarantees and demonstrate the potential strength of hybrid theoretical tools to overcome fundamental limitations of continuous-time networked systems as the non-uniform behavior typically found with classical Kuramoto models.

In a first set of simulations, to illustrate the uniform asymptotic and finite-time synchronization property, we consider a tree-like network of $n = 20$ heterogeneous oscillators with time-varying natural frequency and we initialize solutions far away from the synchronization set, in a neighbourhood of one of the unstable equilibria of the classical Kuramoto oscillator (Strogatz (2000)).

The phase evolution is reported [1] in Fig. 3, for different selections of σ, and $\kappa = 1$. When σ is defined as the quadratic or the sine function, practical synchronization is achieved, as shown in Fig. 3. We note that using the

[1] The simulations have been carried out using the Matlab toolbox HyEQ (Sanfelice et al. (2013)).

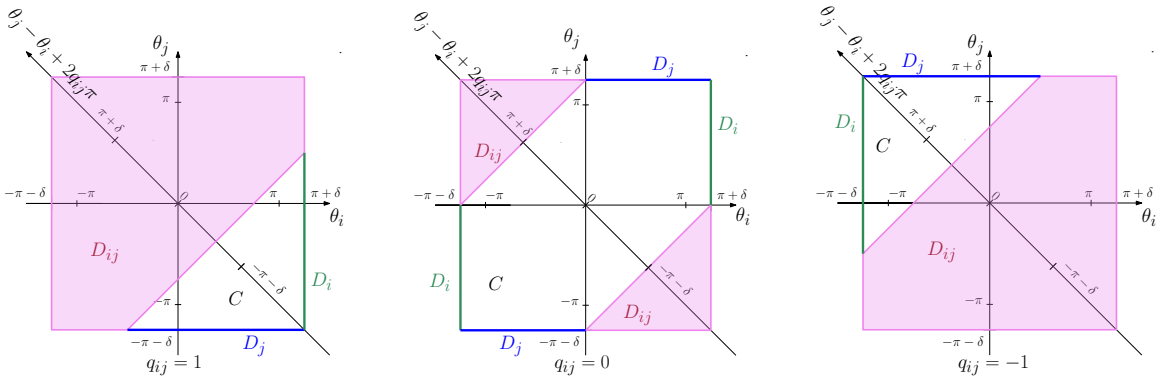


Fig. 1. Projection of the flow and jump sets on (θ_i, θ_j) for each value of q_{ij} (Mariano et al. (2021)).

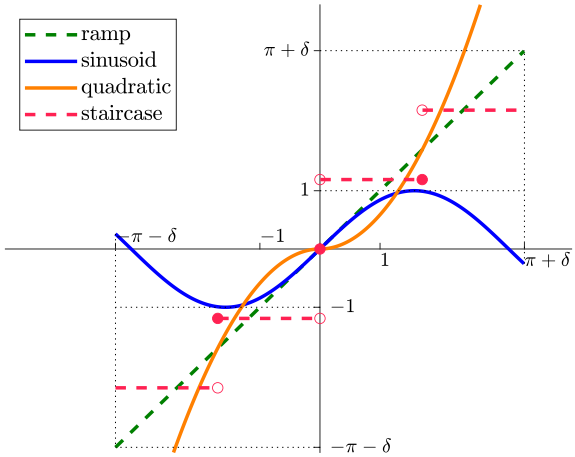


Fig. 2. Examples of suitable functions σ satisfying, together with the sine function used in the classical Kuramoto model.

sinusoidal σ, as in the classical Kuramoto oscillators, leads to a slower convergence due to the non-uniform synchronization property, as compared to the quadratic and the staircase functions. On the other hand, the staircase function, which is discontinuous at 0, leads also to a finite-time synchronization property, in agreement with Theorem 2.

In a second set of simulations, we consider $n = 5$ oscillators, $\kappa = 1$ and $\frac{\kappa}{n} = 1$ in (1), so that the coupling gains are the same and $\omega_i = 1$, $i \in \{1, \ldots, 5\}$. Simulation results for an "all-to-all" network are provided in Fig.4 where we see that the phases generated by our model show a richness of converging behaviors (asymptotic and finite time convergence compared to the exponential one of the classical Kuramoto model) for relatively large initial phase mismatch (left column). For larger initial errors (right column), our designed hybrid coupling induces uniform synchronization, whereas slow transients are generated by (1) (top row). Similar results have been obtained in the case where the interconnection graph is a tree. Even though our main results in Theorems 1 and 2 require tree-like networks, Fig.4 shows that our solution may provide desirable uniform synchronization also with more general "all-to-all" networks.

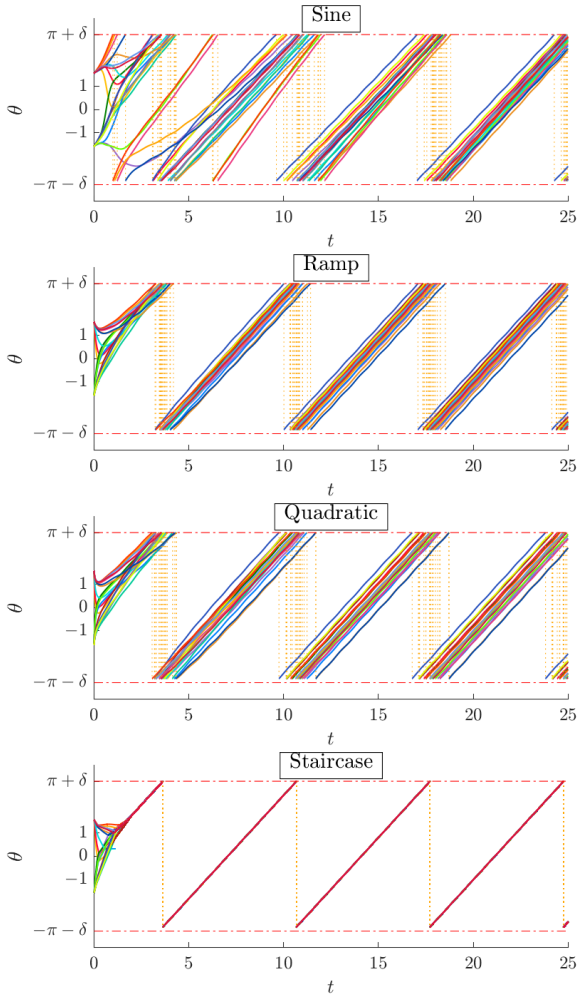


Fig. 3. Phase evolution for $\kappa = 1$ and different selections of σ (from top to bottom: sine, ramp, quadratic and staircase functions).

REFERENCES

Acebron, J.A., Bonilla, L.L., Vicente, C.J.P., Ritort, F., and Spigler, R. (2005). The Kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of Modern Physics*, 77(1), 137–185.

Aeyels, D. and Rogge, J.A. (2004). Existence of partial entrainment and stability of phase locking behavior of coupled oscillators. *Progress of Theoretical Physics*, 112(6), 921–942.

Aokii, T. (2015). Self-organization of a recurrent network under ongoing synaptic plasticity. *Neural Networks*, 62, 11–19.

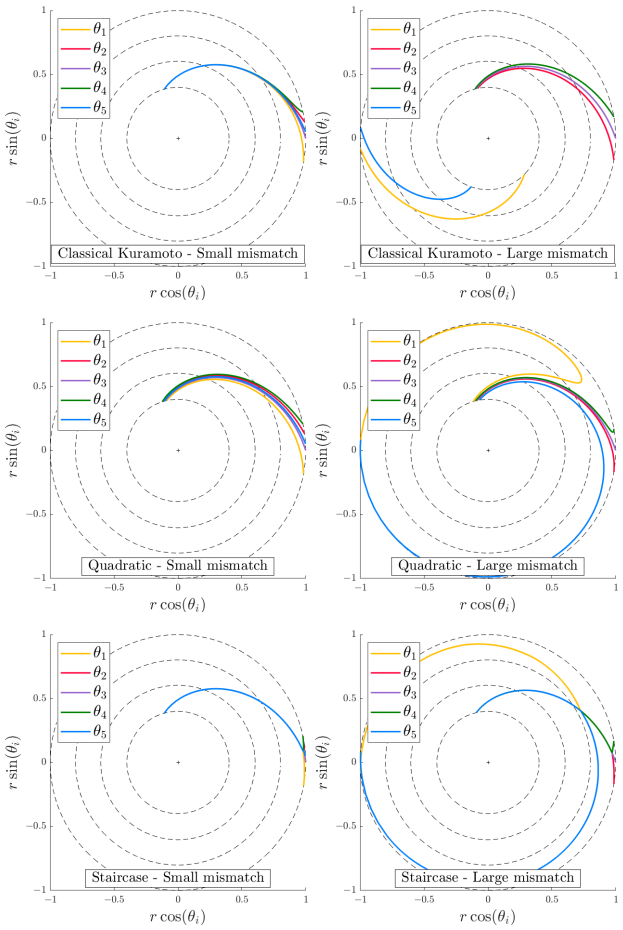


Fig. 4. Evolution of the pair $(r\cos(\theta_i), r\sin(\theta_i))$, with $r(\mathfrak{t}) = -0.2\mathfrak{t}+1$, showing radially the continuous time evolution for the phases generated by the classical Kuramoto model (1) (top) and our hybrid modification (middle-bottom).The initial phase mismatch is increasingly large from left to right. The black dashed lines are isotime (0 (outer), 0.6, 1.2 and 1.8 (inner) time units).

Chopral, N. and Spong, M.W. (2009). On exponential synchronization of Kuramoto oscillators. *IEEE Trans. on Automatic Control*, 54(2), 353–357.

Cumin, D. and Unsworth, C.P.A. (2007). Generalising the Kuramoto model for the study of neuronal synchronisation in the brain. *Physica D: Nonlinear Phenomena*, 226(2), 181–196.

Dörfler, F. and Bullo, F. (2012). Synchronization and transient stability in power networks and nonuniform Kuramoto oscillators. *SIAM Journal on Control and Optimization*, 50(3), 1616–1642.

Dörfler, F. and Bullo, F. (2011). On the critical coupling strength for kuramoto oscillators. In *American Control Conference*, 3239–3244.

Dörfler, F. and Bullo, F. (2014). Synchronization in complex networks of phase oscillators: A survey. *Automatica*, 50(6), 1539–1564.

Forrester, D.M. (2015). Arrays of coupled chemical oscillators. *Scientific Reports*, 5(16994).

Jadbabaie, A., Motee, N., and Barahona, M. (2004). On the stability of the Kuramoto model of coupled nonlinear oscillators. In *American Control Conference*, 4296–4301.

Jafarpour, S. and Bullo, F. (2019). Synchronization of Kuramoto oscillators via cutset projections. *IEEE Trans. on Automatic Control*, 64(7), 2830 – 2844.

Kuramoto, Y. (1975). Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics*, 420–422. Springer.

Leonard, N.E., Shen, T., Nabet, B., Scardovi, L., Couzin, I.D., and Levin, S.A. (2012). Decision versus compromise for animal groups in motion. *Proceedings of the National Academy of Sciences*, 109(1), 227–232.

Mariano, S., Bertollo, R., Postoyan, R., and Zaccarian, L. (2021). Leaderless uniform global asymptotic and finite-time synchronization of Kuramoto-like oscillators. *Submitted for publication to Automatica*.

Miller, R.B. and Pachter, M. (1997). Maneuvering flight control with actuator constraints. *Journal of Guidance, Control, and Dynamics*, 20(4), 729–734.

Oud, W.T. (2006). *Design and experimental results of synchronizing metronomes, inspired by Christiaan Huygens*. Master's Thesis, Eindhoven University of Technology.

Polyakov, A. (2011). Nonlinear feedback design for fixed-time stabilization of linear control systems. *IEEE Trans. on Automatic Control*, 57(8), 2106–2110.

Sanfelice, R.G., Copp, D., and Nanez, P. (2013). A toolbox for simulation of hybrid systems in Matlab/Simulink: Hybrid Equations (HyEQ) Toolbox. In *Proceedings of the 16th International Conference on Hybrid Systems: Computation and Control*, 101–106. ACM.

Sepulchre, R., Paley, D.A., and Leonard, N.E. (2007). Stabilization of planar collective motion: All-to-all communication. *IEEE Trans. on Automatic Control*, 52(5), 811–824.

Song, Y., Wang, Y., Holloway, J., and Krstić, M. (2017). Time-varying feedback for regulation of normal-form nonlinear systems in prescribed finite time. *Automatica*, 83, 243–251.

Strogatz, S.H. (2000). From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1), 1–20.

Tass, P.A. (2003). A model of desynchronizing deep brain stimulation with a demand-controlled coordinated reset of neural subpopulations. *Biol. Cybernetics*, 89(2), 81–88.

Wu, J. and Li, X. (2018). Finite-time and fixed-time synchronization of Kuramoto-oscillator network with multiplex control. *IEEE Trans. on Control of Network Systems*, 6(2), 863–873.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On evolutionary population games on community networks with dynamic densities

A. Govaert * **L. Zino** ** **E. Tegling** *

* *Department of Automatic Control, Lund University, , SE-221 00 Lund, Sweden (email: {alain.govaert,emma.tegling}@control.lth.se)*
** *Faculty of Science and Engineering, University of Groningen, 9747 AG Groningen, The Netherlands (email: lorenzo.zino@rug.nl)*

Abstract: We deal with evolutionary game-theoretic learning processes for population games on networks with dynamically evolving communities. Specifically, we propose a novel framework in which a deterministic, continuous-time replicator equation on a community network is coupled with a closed migration process between the communities, in turn governed by an environmental feedback mechanism resulting in co-evolutionary dynamics. Through a rigorous analysis of the system of differential equations obtained, we characterize the equilibria of the coupled dynamical system. Moreover, for a class of population games —matrix games— a Lyapunov argument is used to establish an evolutionary folk theorem that guarantees convergence to the evolutionary stable states of the game. Numerical simulations are provided to illustrate and corroborate our theoretical findings.

Keywords: Game theory, evolutionary games, flows in graphs.

1. INTRODUCTION

The literature on evolutionary game theory usually relies on the assumption that individuals interact on a homogeneous time-invariant all-to-all communication structure. However, this assumption is quite simplistic in many real-world scenarios Easley and Kleinberg (2010). To address this limitation, in particular for the class of learning mechanisms regulated by pairwise interactions and imitation dynamics, some recent efforts toward incorporating a mesoscopic network structure into learning protocols have been made. In these frameworks, it is assumed that the players are divided into communities that determine their possible interactions with other players. For example, in Hofbauer and Sandholm (2009); Sandholm (2010), communities are introduced as fully mixed and isolated populations, where players do not interact across communities, but the communities themselves are coupled through a common payoff function. For such a model, global convergence results have been established for several classes of games. In Barreiro-Gomez et al. (2017), some convergence results have been extended to populations in which the the action played by an individual determines their community and thus, ultimately, their pattern of interactions. In Como et al. (2021), imitation dynamics on community networks have been formalized, in which players belong to different communities who interact on a network structure. Some convergence results have been established, including global convergence for potential games.

[1] This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

In the aforementioned works, it is assumed that the communities are fixed a priori Como et al. (2021); Hofbauer and Sandholm (2009); Sandholm (2010) or determined by the individuals' actions Barreiro-Gomez et al. (2017). This relies on an assumption of time-scale separation, in which a dynamical co-evolution of the communities at the same time-scale as the learning process is neglected.

In this work, we address this gap by proposing a novel dynamic coupling of two mechanisms. On the one hand, evolutionary dynamics on community networks with closed migration processes as in Kelly (2011).On the other, environmental feedback, which has previously been modeled for evolutionary game frameworks *without* community structure in Tilman et al. (2020). Here, we model a scenario where individuals of a community can move to other communities in response to environmental changes. This means that we augment the system of ordinary differential equations (ODEs) that characterizes the replicator equation on community networks from Como et al. (2021) with a set of ODEs that describes the evolution of the community densities.

Notation: The sets of real and non-negative real numbers are denoted by $\mathbb{R}$ and $\mathbb{R}_+$, respectively. For finite sets $\mathcal{A}$ and $\mathcal{B}$, $\mathbb{R}^{\mathcal{A}\times\mathcal{B}}$ denotes the set of real matrices whose entries are indexed by the elements of $\mathcal{A}\times\mathcal{B}$. The transpose of a matrix $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^\top$. The j-th column (row) of matrix $\boldsymbol{x}$ is denoted by $\boldsymbol{x}_j$ ($\boldsymbol{x}_{j'}$) and the ij-th element by x_{ij}. The i-th element of a vector $\boldsymbol{y}$ is denoted by y_i and the 2-norm of the vector as $||\boldsymbol{y}||$ The vector of all ones is denoted by $\mathbf{1}$ and the sign function is denoted by sgn. For a non-negative matrix W in $\mathbb{R}^{n\times n}$ the associated graph is

defined as $(\mathcal{N}, \mathcal{E}_W)$, with node set $\mathcal{N} := \{1, 2, \dots, n\}$ and edge set $\mathcal{E}_W := \{(i,j) \in \mathcal{N} \times \mathcal{N} : W_{ij} > 0\}$.

2. MODEL

We consider a continuum of individuals structured into communities that interact with each other through instantaneous, random, pairwise encounters with varying strengths, both within and between the communities in the population. In each pairwise encounter individuals use an action from a finite and common action set, which, together with the action of the opponent, results in a reward. Evolutionary dynamics describe how the frequency of actions change under the influence of the pairwise encounters. Here, we consider a replicator equation on community networks in which the frequency of actions is assumed to be proportional to the expected reward and performance of the actions in the population. The novel aspect is that individuals can move freely between the communities. The communities are connected by a dynamic flow network whose flow rates change in response to the frequencies of actions in the communities and, possibly, an exogenous process. Since the movement of individuals changes the rate at with which pairwise encounters occur, a feedback process is established that describes a coevolutionary process of strategic interaction and migration at a community and population level. In the following, the formal definitions of the various concepts are provided.

2.1 Population game

Given a finite action set the *population state* $\boldsymbol{y}$ is a vector in the unitary simplex over $\mathcal{A}$ defined as $\mathcal{Y} := \{\boldsymbol{y} \in \mathbb{R}_+^{\mathcal{A}} : \mathbf{1}^\top \boldsymbol{y} = 1\}$. The elements y_i of the population state $\boldsymbol{y}$ denotes the fraction of players in the population that use action i in $\mathcal{A}$ (i-players). A population state $\boldsymbol{y}$ in $\mathcal{Y}$ is said to support action i in $\mathcal{A}$ if a non-zero fraction of the population uses it. The set of $\mathcal{S}_{\boldsymbol{y}} := \{i \in \mathcal{A} : y_i > 0\}$ is called the *support* of $\boldsymbol{y}$. Given a population state expected rewards $r_i(\boldsymbol{y})$ are determined by the reward functions $r_i : \mathcal{Y} \to \mathbb{R}$ for i in $\mathcal{A}$. A *population game* then refers to the pair $(\mathcal{Y}, r)$.

2.2 Community network

Individuals are structured into a finite set $\mathcal{H}$ of communities. We refer to the proportion of the population in community h in $\mathcal{H}$ as the *community density* and denote it by η_h. The fraction of i-players in community h is denoted by x_{ih} and make up the elements of the *system state* matrix $\boldsymbol{x}$ in $\mathbb{R}_+^{\mathcal{A}\times\mathcal{H}}$. The columns of the system state matrix are referred to as the *community state* vectors $\boldsymbol{x}_h$ in $\mathbb{R}_+^{\mathcal{A}}$ for h in $\mathcal{H}$. Similar as before, the support of a community state is $\mathcal{S}_{\boldsymbol{x}_h} := \{i \in \mathcal{A} : x_{ih} > 0\}$ and $\cup_{h\in\mathcal{H}} \mathcal{S}_{\boldsymbol{x}_h} = \mathcal{S}_{\boldsymbol{y}}$. The density of the population is assumed to be constant and given by

$$\boldsymbol{\eta}\mathbf{1} = \mathbf{1}^\top \boldsymbol{x}\mathbf{1} = \mathbf{1}^\top \boldsymbol{y} = 1, \tag{1}$$

which results in the set of admissible system states $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^{\mathcal{A}\times\mathcal{H}} : (1)\}$. The strength of interactions between communities is determined by a constant non-negative matrix W in $\mathbb{R}_+^{\mathcal{H}\times\mathcal{H}}$. Together with the fraction of i-players this determines the rate $x_{ih}W_{hk}x_{jk} \geq 0$, for i, j in $\mathcal{A}$ and h, k in $\mathcal{H}$, at which i-players in community h meet j-players in community k in a pairwise encounter. We refer to the triplet $(\mathcal{H}, W, \boldsymbol{\eta})$ as a community network. Throughout the extended abstract the following assumption is made.

Assumption 1. W is non-negative and irreducible with a strictly positive diagonal. That is, the graph $(\mathcal{H}, \mathcal{E}_W)$, associated to the community network is connected and has self-loops.

2.3 Evolutionary dynamics

Although the results that we will describe in section 3.1 can be generalized to a broader class of evolutionary imitation dynamics, here we focus on the replicator equation due to its prominence in evolutionary game theory Cressman et al. (2003); Cressman and Tao (2014) and control applications of population games Quijano et al. (2017). The replicator equation on a community network is a matrix valued equation $\boldsymbol{f}(\boldsymbol{x})$ in $\mathbb{R}_+^{\mathcal{H}\times\mathcal{H}}$ whose elements Como et al. (2021)

$$f_{ih}(\boldsymbol{x}) = \eta_h \sum_{k\in\mathcal{H}} x_{ik}W_{hk}r_i(\boldsymbol{y}) - x_{ih}\sum_{j\in\mathcal{A}}\sum_{k\in\mathcal{H}} x_{jk}W_{hk}r_j(\boldsymbol{y}). \tag{2}$$

describe how the proportion of i-players in community h changes under the influence of selection. We remark that, if there is a community h in $\mathcal{H}$ such that $\eta_h = 1$, then (2) reduces to the more familiar form of the classic replicator equation Cressman and Tao (2014) given by $f_i(\boldsymbol{x}) = x_i\left(r_i(\boldsymbol{x}) - \sum_{j\in\mathcal{A}} x_j r_j(\boldsymbol{x})\right)$ for $x_i = W_{hh}x_{ih}$.

Assumption 2. Expected rewards are positive.

This assumption is born out of a technical necessity that is not confining: because the restricted Nash equilibria of (2) are invariant to the addition of a constant, negative rewards can always be changed to positive ones without changing the set of equilibrium points.

2.4 Dynamic flow process

We assume individuals of a community have an intrinsic tendency for movement that is described by a constant non-negative matrix Λ in $\mathbb{R}_+^{\mathcal{H}\times\mathcal{H}}$ with elements λ_{hk}. Given a system state $\boldsymbol{x}$ in $\mathcal{X}$ these intrinsic tendencies may be amplified or reduced by a non-negative environmental function. $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}_+^{\mathcal{H}\times\mathcal{H}}$. The environmental response function may also depend on a subset of community state vectors or an exogenous variable. It may also be governed by positive system dynamics. As in the closed migration processes of (Kelly, 2011, Chapter 2), we assume scaling is multiplicative such that the dynamic rate at which individuals move from community h to community k is $\lambda_{hk}\phi_{hk}(\boldsymbol{x})$ in $\mathbb{R}_+$. The changes in community densities induced by these movements are described by the dynamic flow process

$$\dot{\eta}_h = \sum_{k\in\mathcal{H}} \lambda_{kh}\phi_{kh}(\boldsymbol{x})\eta_k - \eta_h \sum_{k\in\mathcal{H}} \lambda_{hk}\phi_{hk}(\boldsymbol{x}). \tag{3}$$

This preserves the population density (1) because the system is closed. Moreover, non-negativity of the environmental function $\boldsymbol{\phi}$ and movement matrix Λ ensures the solutions of (3) remain well-defined densities.

This can, for example, model density-limiting effects, which are an important consideration in ecological and population models. Increased pressures on resources can lower reproductive rates Cressman and Garay (2003) and increase out-migration Isard (1960); Masanori et al. (2015). The latter can, for example, be captured by considering a dynamic environmental response that is uniform in the *outflows* of a community:

$$\dot{\phi}_{hk} = (\phi_{hk} - m)\left(1 - \frac{\eta_h}{\kappa_h}\right)\phi_{hk}, \tag{4}$$

where $m > 0$ is a constant maximum environmental response and $\kappa_h > 0$ the carrying capacity of community h. Clearly, its solutions $\phi_h(t)$ of the above differential equation exist in $\mathbb{R}_+$ as required for well-defined community densities. Moreover, if a community is overcrowded $\eta_h > \kappa_h$ and out-migration increases. This example may be generalized to account for a dependency of the carrying capacity on the community state vector x_h akin to density games Novak et al. (2013).

When the dynamic movement rates are frequency dependent, but action independent, the proportion of actions in the outflows of a community are distributed according to the corresponding community state vector. Thus, when (2) is interconnected with (3), the closed-loop system state dynamics read as

$$\begin{aligned}\dot{x}_{ih} &= \sum_{k\in\mathcal{H}} \lambda_{kh}\phi_{kh}(\boldsymbol{x})x_{ik} - x_{ih}\sum_{k\in\mathcal{H}} \lambda_{hk}\phi_h(\boldsymbol{x}) \\ &+ \eta_h \sum_{k\in\mathcal{H}} x_{ik}W_{hk}r_i(\boldsymbol{y}) - x_{ih}\sum_{j\in\mathcal{A}}\sum_{k\in\mathcal{H}} x_{jk}W_{hk}r_j(\boldsymbol{y}).\end{aligned} \tag{5}$$

Definition 1. The combination of a population game, a community network and an environmental function, define the population game on a community network with dynamic densities as the tuple $\Gamma = (\mathcal{Y}, r, \mathcal{H}, W, \Lambda, \boldsymbol{\phi})$.

3. RESULTS

3.1 A dynamic system state-density balance

We next characterize the asymptotic relation between the population state, the community state and dynamic community densities of a connected community network. For our first result, no further restrictions are imposed on the size of the finite action set, or the structure of the reward functions and environmental function.

Theorem 1. Consider a population game on a community network with dynamic densities Γ that satisfies Assumptions 1-2. Let $\boldsymbol{\eta}(t)$ and $\boldsymbol{x}(t)$ be the solutions of the dynamics (3) and (5), respectively. If $\lim_{t\to\infty} \boldsymbol{x}(t)\mathbf{1} = \boldsymbol{y}^*$, then $\boldsymbol{y}^*$ is a restricted Nash equilibrium and, for all h in $\mathcal{H}$ such that $\liminf_{t\to\infty} \eta_h(t) > 0$, it holds that

$$\lim_{t\to\infty} \frac{x_{ih}(t)}{\eta_h(t)} = y_i^* \quad \forall i \in \mathcal{A}. \tag{6}$$

In the degenerate cases $\eta_k(t) = 0$, $x_{ik}(t) = 0$ for all i in $\mathcal{A}$.

The proof is based on W being non-negative and irreducible, together with the dynamic flow process being closed and non-negative. Using this, it can be shown that the ratio x_{ih}/η_h is constant and equal for all communities if the population state $\boldsymbol{y}$ is at an equilibrium. By Assumption 2, this must be a restricted Nash equilibrium. Details can be found in Govaert et al. (2022).

Theorem 1 shows that the dynamic system state-density balance in(6) is achieved even when the dynamic flow process (3) is non-convergent, i.e. $\lim_{t\to\infty} \boldsymbol{\eta}(t)$ does not exist. Otherwise, the following corollary is an immediate consequence of Theorem 1.

Corollary 1.1. If the dynamic flow process (3) converges to $\lim_{t\to\infty} \boldsymbol{\eta}(t) = \boldsymbol{\eta}^*$ and $\lim_{t\to\infty} \boldsymbol{x}(t)\mathbf{1} = \boldsymbol{y}^*$, then the system state matrix converges to the equilibrium $\lim_{t\to\infty} \boldsymbol{x}(t) = \boldsymbol{y}^*\boldsymbol{\eta}^{*\top}$.

3.2 Evolutionary stability

Next, we focus on the relation between the population state vector and evolutionarily stable states. For this result, we restrict our attention to binary action sets and linear rewards functions of the form

$$r(\boldsymbol{y}) = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\boldsymbol{y} \quad a, b, c, d > 0, \tag{7}$$

that satisfy Assumption 2. These rewards can also be interpreted as the payoffs of a player in a two-by-two symmetric matrix game played against the mixed strategy $\boldsymbol{y}$. Consider the following definition of an evolutionarily stable state from Hofbauer and Sigmund (1998).

Definition 2. (Evolutionarily stable state). A population state vector $\hat{\boldsymbol{y}}$ in $\mathcal{Y}$ is an evolutionarily stable state if there exists $\delta > 0$ such that $\hat{\boldsymbol{y}} \cdot A\boldsymbol{y} > \boldsymbol{y} \cdot A\boldsymbol{y}$, for all $\boldsymbol{y} : 0 < ||\boldsymbol{y} - \hat{\boldsymbol{y}}|| < \delta]$.

The following result shows the importance of evolutionarily stable states also for the replicator equation on networks with dynamic communities when the underlying interaction network is undirected. Its proof, omitted due to space constraints, is based on a Lyapunov argument.

Theorem 2. Consider Γ that satisfies Assumption 1 and additionally $W = W^\top$. The action set is binary and reward functions are linear and symmetric. Then,

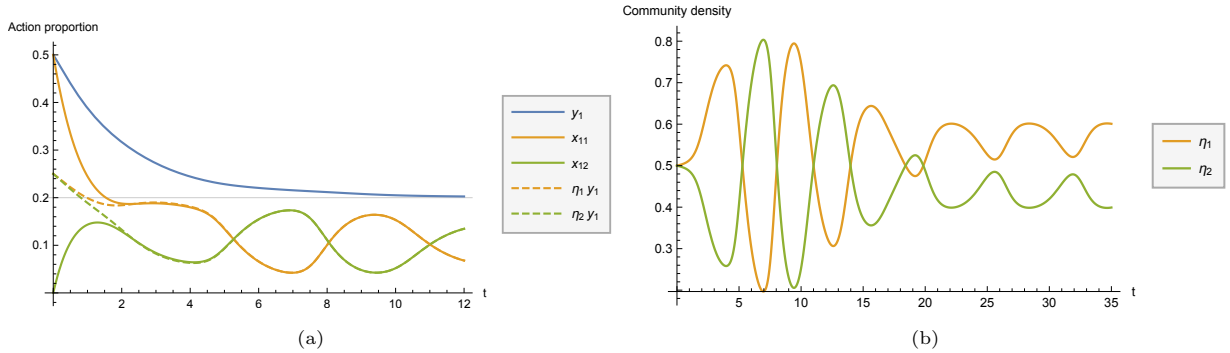
(1) An evolutionarily stable state $\hat{\boldsymbol{y}}$ in $\mathcal{Y}$ is locally asymptotically stable.
(2) If an evolutionarily stable state $\hat{\boldsymbol{y}}$ exists in the interior of $\mathcal{Y}$ then all interior trajectories converge to it.

The proof is based on a Lyapunov argument, where the strict local Lyapunov function $P(\boldsymbol{y}) := \prod_{i\in\mathcal{S}_{\hat{y}}} y_i^{\hat{y}_i}$ is combined with the assumption that W is connected and symmetric. Details can be found in Govaert et al. (2022).

The combination of Theorems 1 and 2 characterizes the asymptotic behavior at a population and community level and shows the important role of evolutionarily stable states. We now illustrate this with a brief case study on carrying capacities that exhibit periodic behaviors under the influence of geographically distributed seasonal changes or cyclic socio-economic parameters. We illustrate this for two communities with a sinusoidally varying carrying capacities Banks (1993):

$$\kappa_1(t) = \gamma\,\sin(t) + \rho, \quad \kappa_2(t) = \gamma\,\sin(t + \pi) + \rho, \tag{8}$$

with $0 < \gamma < \rho$ to ensure they are positive. The phase shift between the two communities represents a geographical



(a) (b)

Fig. 1. Numerical solutions of the dynamic flow process (3) and the closed-loop system (5) for two communities with the dynamic environmental function (4) with sinusoidal varying carrying capacities (8). Notice in (b) that the community densities oscillate periodically from $t = 22$, while the population state in (a) converges to an equilibrium and a dynamic balance system state is achieved asymptotically. Parameters values are as follows. Carrying capacity: $\gamma = 0.25$, $\rho = 0.5$. Pairwise rewards: $a = 1$, $b = 7$, $c = 5$, $d = 6$. Interaction matrix: $W_{11} = 0.7$, $W_{12} = W_{21} = W_{22} = 0.3$. Movement matrix: $\lambda_{aa} = \lambda_{bb} = 1$, $\lambda_{ba} = 0.8$, $\lambda_{ab} = 0.5$. Initial conditions not shown in the figure: $\phi_{12}(0) = \phi_{21}(0) = 0.05$.

difference in seasonal changes or socio-economic parameters. The sinusoidally varying carrying capacities can be combined with the dynamic environmental function (4). A closed system of differential equations is then obtained with the dynamic flow process (3) and the evolutionary dynamic (5) with rewards (7). Even for just two communities a full analysis is challenging. However, with the theory developed here, some critical insights at both a population and community level can be obtained. As Theorem 2 predicts, the population state converges asymptotically to the evolutionarily stable state indicated by the horizontal line in Fig. 1a. The interesting behavior occurs at a community level that shows persistent oscillations due to the sinusoidally varying carrying capacities as in Fig. 1b. The effect of Theorem 1 is then seen by the trajectories that converge to each other in Fig. 1a: the proportion of players in the communities converge asymptotically to the product of the oscillating community densities and the evolutionarily stable state.

4. CONCLUSION

In this work, we have proposed a novel framework for evolutionary game dynamics on networks with dynamic communities. Specifically, our framework couples a learning dynamics on a community network with closed migration processes and an environmental feedback, which coevolve at comparable time scales. Under some reasonable assumptions on the structure of the networks and on the reward functions, we have provided a characterization of the equilibria of the dynamical system. Moreover, for matrix games on undirected networks, we have established a convergence result to the evolutionary stable states of the system.

REFERENCES

Banks, R.B. (1993). *Growth and diffusion phenomena: Mathematical frameworks and applications*, volume 14. Springer Science & Business Media.

Barreiro-Gomez, J., Obando, G., and Quijano, N. (2017). Distributed population dynamics: Optimization and control applications. *IEEE Trans. Syst., Man, Cybern. Syst.*, 47(2), 304–314.

Como, G., Fagnani, F., and Zino, L. (2021). Imitation dynamics in population games on community networks. *IEEE Trans. Control. Netw.*, 8(1), 65–76.

Cressman, R., Ansell, C., and Binmore, K. (2003). *Evolutionary dynamics and extensive form games*, volume 5. MIT Press.

Cressman, R. and Garay, J. (2003). Stability in n-species coevolutionary systems. *Theor. Popul. Biol.*, 64(4), 519–533.

Cressman, R. and Tao, Y. (2014). The replicator equation and other game dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 111(Supplement 3), 10810–10817.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: reasoning about a highly connected world*. Cambridge University Press.

Govaert, A., Zino, L., and Tegling, E. (2022). Population games on dynamic community networks. *IEEE Contr. Syst. Lett.*, 6, 2695–2700.

Hofbauer, J. and Sandholm, W.H. (2009). Stable games and their dynamics. *J. Econ. Theory*, 144(4), 1665 – 1693.e4.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Isard, W. (1960). *Methods of regional analysis*. MIT Press.

Kelly, F.P. (2011). *Reversibility and Stochastic Networks*. Cambridge University Press.

Masanori, T., Wada, K., and Fukuda, I. (2015). Environmentally driven migration in a social network game. *Scientific Reports*, 5(12481).

Novak, S., Chatterjee, K., and Nowak, M.A. (2013). Density games. *J. Theor. Biol.*, 334, 26–34.

Quijano, N., Ocampo-Martinez, C., Barreiro-Gomez, J., Obando, G., Pantoja, A., and Mojica-Nava, E. (2017). The role of population games and evolutionary dynamics in distributed control systems: The advantages of evolutionary game theory. *IEEE Contr. Syst. Mag.*, 37(1), 70–97.

Sandholm, W.H. (2010). *Population Games and Evolutionary Dynamics*. Cambridge University Press.

Tilman, A.R., Plotkin, J.B., and Akçay, E. (2020). Evolutionary games with environmental feedbacks. *Nat. Commun.*, 11(915).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Schrödinger Bridges on Trees and Multi-Marginal Optimal Transport ⋆

Isabel Haasler * **Axel Ringh** ** **Yongxin Chen** ***
Johan Karlsson *

* *KTH Royal Institute of Technology, Stockholm, Sweden (e-mail: haasler@kth.se, johan.karlsson@math.kth.se).*
** *Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden (e-mail: axelri@chalmers.se).*
*** *Georgia Institute of Technology, Atlanta, GA, USA (e-mail: yongchen@gatech.edu).*

Abstract: Recently, there has been a large interest in the theory of optimal transport and its connections to the Schrödinger bridge problem. In this work we generalize some of these results to multi-marginal optimal transport problems when the cost function decouples according to a tree structure. In particular, the entropy regularized multi-marginal optimal transport problem can be seen as a Schrödinger bridge problem on the same tree. Moreover, based on this, we extend efficient algorithms for the bi-marginal problem to the multi-marginal setting where the cost function decouples according to a tree structure. Such problems appear in several applications of interest such as barycenter and tracking problems. A common approach for solving these problems is by utilizing pairwise regularization. However, we show that the multi-marginal regularization introduces less diffusion, which is favorable in many applications.

Keywords: Optimal transport, Schrödinger bridge, Graph signal processing, Multi-marginal problems, Entropy regularization

1. MULTI-MARGINAL OPTIMAL TRANSPORT

An optimal transport problem is to find a transport plan that minimizes the cost of moving mass from one distribution to another, see, e.g., Villani (2008). Historically this problem has been important in economics and operations research, but as a result of recent progress it has become a popular tool in a wide range of fields such as signal processing, computer vision, automatic control, and machine learning (see, e.g., Elvander et al. (2020); Dominitz and Tannenbaum (2010); Solomon et al. (2015); Chen et al. (2016c); Adler et al. (2017)). An extension to the standard optimal transport framework is multi-marginal optimal transport, see, e.g., Pass (2015), which seeks a transport plan between not only two, but several distributions.

In this extended abstract, which is based on Haasler et al. (2021), we consider discrete multi-marginal optimal transport problems. In this setting, the marginal distributions are described by nonnegative vectors $\mu_j \in \mathbb{R}^n_+$, for $j = 1, \dots, J$, and we seek a nonnegative J-mode mass transport tensor $\mathbf{M} \in \mathbb{R}^{n\times n\cdots\times n}_+$ that minimizes the transportation cost between the marginals. In particular, the element $\mathbf{M}_{i_1,\dots,i_J}$ of $\mathbf{M}$ denotes the amount of transported mass associated with the tuple $(i_1, \dots, i_J)$, where $i_j \in \{1, \dots, n\}$ describes the location on the j-th marginal, and the tensor $\mathbf{M}$ is thus a transport plan between the marginals $\mu_j \in \mathbb{R}^n_+$, for $j = 1, \dots, J$, if its projections satisfy $P_j(\mathbf{M}) = \mu_j$, for $j = 1, \dots, J$, where

$$P_j(\mathbf{M}) = \sum_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_J} \mathbf{M}_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_J}.$$

The cost for transporting mass is described by the nonnegative J-mode cost tensor $\mathbf{C} \in \mathbb{R}^{n\times n\cdots\times n}_+$, where $\mathbf{C}_{i_1,\dots,i_J}$ assigns a cost to the tuple $(i_1, \dots, i_J)$. The total cost is thus $\langle \mathbf{C}, \mathbf{M}\rangle = \sum_{i_1,\dots,i_J} \mathbf{C}_{i_1,\dots,i_J}\mathbf{M}_{i_1,\dots,i_J}$, and the problem of finding an optimal transport plan $\mathbf{M}$ can therefore be formulated as a linear program. However, in many practical applications this linear program is impossible to solve directly due to the large number of variables, n^J. In Cuturi (2013) these computational limitations have been alleviated for the bi-marginal setting by introducing an entropy regularization. Following the same approach, we formulate the entropy regularized multi-marginal optimal transport problem (see also Benamou et al. (2015); Elvander et al. (2020)) as

$$\begin{aligned} &\underset{\mathbf{M}\in\mathbb{R}^{n\times\cdots\times n}_+}{\text{minimize}} \ \langle \mathbf{C}, \mathbf{M}\rangle + \epsilon H(\mathbf{M}) \\ &\text{subject to } P_j(\mathbf{M}) = \mu_j, \ \text{for } j \in \Gamma, \end{aligned} \tag{1}$$

where $\epsilon > 0$ is a regularization parameter, the normalized entropy of $\mathbf{M}$ is defined as

$$H(\mathbf{M}) := \sum_{i_1,\dots,i_J} \left(\mathbf{M}_{i_1,\dots,i_J} \log\left(\mathbf{M}_{i_1,\dots,i_J}\right) - \mathbf{M}_{i_1,\dots,i_J} + 1\right),$$

and $\Gamma \subset \{1, 2, \dots, J\}$ is an index set. At this point, note in particular that the latter means that not all marginal projections of $\mathbf{M}$ need to be assigned, as in the Wasserstein barycenter problem. It can be shown that the optimal solution to (1) is of the form $\mathbf{M} = \mathbf{K} \odot \mathbf{U}$, where $\odot$ denotes elementwise multiplication, $\mathbf{K} = \exp(-\mathbf{C}/\epsilon)$ and $\mathbf{U}$ can be

⋆ This work was supported by the Swedish Research Council (VR), grant 2014-5870, SJTU-KTH cooperation grant and the NSF under grant 1901599. This work was previously accepted to MTNS 2020.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

decomposed as

$$\mathbf{U} = u_1 \otimes u_2 \otimes \cdots \otimes u_J \tag{2}$$

with $u_j \in \mathbb{R}^n$ given by

$$u_j = \begin{cases} \exp(\lambda_j/\epsilon), & \text{if } j \in \Gamma \\ \mathbf{1}, & \text{else.} \end{cases}$$

Here $\lambda_j \in \mathbb{R}^n$ is the optimal dual variable corresponding to the constraint on the j-th marginal in the dual problem of (1). For details the reader is referred to, e.g., Elvander et al. (2020); Benamou et al. (2015).

The optimal $\mathbf{U}$ in (2) can be found by so-called Sinkhorn iterations, which are given by iteratively updating u_j according to

$$u_j \leftarrow u_j \odot \mu_j ./ P_j(\mathbf{K} \odot \mathbf{U}), \tag{3}$$

for all $j \in \Gamma$, where $./$ denotes elementwise division. In Benamou et al. (2015), this scheme was derived as Bregman projections, and in Elvander et al. (2020) as a block coordinate ascend in the dual. If an optimal solution to (1) exists, i.e., if the linear program is feasible, then the Sinkhorn iterations in (3) converge. The computational bottleneck of the Sinkhorn iterations (3) is computing the projections $P_j(\mathbf{M})$, for $j \in \Gamma$, which in general scales exponentially in J. In fact, even storing the tensor $\mathbf{M}$ is a challenge as it consists of n^J elements. However, in many cases of interest, structures in the cost tensors can be exploited to make the computation of the projections feasible, e.g., for computing Euler flows (see Benamou et al. (2015)) and in tracking and information fusion applications (see Elvander et al. (2020)).

2. MULTI-MARGINAL OPTIMAL TRANSPORT ON A TREE

In this section we generalize the methods for solving structured multi-marginal optimal transport problems in Elvander et al. (2020) to the setting where the cost tensor decouples according to a tree structure, that is, when the marginals of the optimal transport problem are associated with the nodes of a tree, and cost matrices are defined on its edges. Such structures appear in various applications of optimal transport. For instance, path trees are often used in tracking and interpolation applications, e.g., in Chen and Karlsson (2018); Solomon et al. (2015). Similarly, star trees describe barycenter problems, which occur for instance in information fusion applications (see, e.g., Cuturi and Doucet (2014); Elvander et al. (2019); Solomon et al. (2015)).

Definition 1. A graph $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, with vertices $\mathcal{V}$ and edges $\mathcal{E}$, is a tree if it is acyclic and connected. The vertices with degree 1 are called leaves; the set of leaves is denoted $\mathcal{L}$. For a vertex $j \in \mathcal{V}$, the set of neighbours $\mathcal{N}_j$ is defined as the set of vertices, which have a common edge with j.

Let $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ be a tree. Consider the entropy regularized multi-marginal optimal transport problem (1) where the marginals correspond to the nodes of the tree, i.e., $\mathcal{V} = \{1, 2, \ldots, J\}$. Assume that the cost tensor $\mathbf{C} \in \mathbb{R}_+^{n^J}$ decouples as

$$\mathbf{C}_{i_1,\ldots,i_J} = \sum_{(j_1,j_2)\in\mathcal{E}} C^{(j_1,j_2)}_{i_{j_1},i_{j_2}}, \tag{4}$$

where $C^{(j_1,j_2)} \in \mathbb{R}^{n\times n}$ is a cost matrix representing the cost of moving mass between marginals μ_{j_1} and μ_{j_2}, for $(j_1, j_2) \in \mathcal{E}$. Note that the cost on edge $(j_1, j_2) \in \mathcal{E}$ can be interchangeably expressed by $C^{(j_2,j_1)}$ without changing the cost tensor $\mathbf{C}$ in (4) by letting $C^{(j_2,j_1)} = (C^{(j_1,j_2)})^T$.

Given: Tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ with leaves $\mathcal{L} = \{1, 2, \ldots, |\mathcal{L}|\}$.
Initial guess u_j, for $j \in \mathcal{L}$
for $(j_1, j_2) \in \mathcal{E}$ **do**
  Initialize $\alpha_{(j_1,j_2)}$ according to (5)
end for
Initialize $j \in \mathcal{L}$
while Sinkhorn not converged **do**
  $u_j \leftarrow u_j \odot \mu_j ./ \bigodot_{k\in\mathcal{N}_j} \alpha_{(j,k)}$
  for $(j_1, j_2) \in \mathcal{E}$ on the path from j to $(j + 1 \mod |\mathcal{L}|)$ **do**
    Update $\alpha_{(j_1,j_2)}$ according to (5)
  end for
  $j \leftarrow j + 1 \mod |\mathcal{L}|$
end while
return u_j for $j \in \mathcal{L}$

Algorithm 1. Sinkhorn method for the multi-marginal optimal transport problem on a tree.

The following theorem describes how to compute the projections of a tensor of the form $\mathbf{K} \odot \mathbf{U}$, where $\mathbf{U} = u_1 \otimes \cdots \otimes u_j$ and $\mathbf{K}$ decouples according to a tree, which appear in the Sinkhorn iterations (3) for optimal transport problems with tree-structured costs.

Theorem 1. Let $\mathbf{K} = \exp(-\mathbf{C}/\epsilon)$ with $\mathbf{C}$ as in (4) for the tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, and let $\mathbf{U} = u_1 \otimes u_2 \otimes \cdots \otimes u_J$. Define $K^{(j_1,j_2)} = \exp(-C^{(j_1,j_2)}/\epsilon)$, for $(j_1, j_2) \in \mathcal{E}$. Then the projection on the j-th marginal of $\mathbf{K} \odot \mathbf{U}$ is of the form

$$P_j(\mathbf{K} \odot \mathbf{U}) = u_j \odot \bigodot_{k\in\mathcal{N}_j} \alpha_{(j,k)}.$$

The vectors $\alpha_{(j,k)}$, for all ordered tuples $(j, k) \in \mathcal{E}$, can be computed recursively starting in the leaves of the tree according to

$$\begin{aligned} \alpha_{(j,k)} &= K^{(j,k)} u_k && \text{for } k \in \mathcal{L} \\ \alpha_{(j,k)} &= K^{(j,k)} \Big(u_k \odot \bigodot_{\ell\in\mathcal{N}_k\setminus\{j\}} \alpha_{(k,\ell)} \Big) && \text{for } k \notin \mathcal{L}. \end{aligned} \tag{5}$$

The expressions for the projections in Theorem 1 can be used to solve a multi-marginal optimal transport problem on a tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ by a Sinkhorn method as detailed in (3). Such a method is summarized in Algorithm 1, where we have also used the observation that in each iteration step some of the factors $\alpha_{(j_1,j_2)}$ do not change. More precisely, between two consecutive updates of u_{j_1} and u_{j_2} only the factors on all edges that lie on the path between nodes j_1 and j_2 are changed and thus need to be updated.

It should be noted that we restrict ourselves to the case where $\Gamma = \mathcal{L}$, since otherwise one can always formulate a multi-marginal optimal transport problem on a subtree (or a set of subtrees), where the marginals are known exactly on the set of leaves of the subtree, and which fully describes the solution to the original problem (see Haasler et al. (2021) for more details).

3. CONNECTIONS TO SCHRÖDINGER BRIDGES

Schrödinger (1931) studied the problem of determining the most likely evolution of a particle cloud observed at two

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

time instances, where the particle dynamics have deviated from the expected Brownian motion. This problem is tightly connected to the classical bi-marginal optimal transport problem, which has been extensively studied in Chen et al. (2016b,a); Léonard (2014). In a discrete setting, the Schrödinger bridge problem can be formulated by modeling the evolutions of a number of particles as a Markov chain (see Pavon and Ticozzi (2010); Georgiou and Pavon (2015)). In Haasler et al. (2019) a similar problem has been considered for hidden Markov chains. In this work, we extend the framework to Markov processes on arbitrary tree-structures, and show that its solution is equivalent to the solution of a certain entropy regularized multi-marginal optimal transport problem on the same tree.

To this end, given a tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ we consider a rooted, directed tree $\mathcal{T}_r(\mathcal{V}, \mathcal{E}_r)$, rooted in a leaf $r \in \mathcal{L}$. The directions of the edges $\mathcal{E}_r$ are uniqely specificed by the condition that any node $j \in \mathcal{V}$ is reachable from the root r. Moreover, this introduces a partial ordering on $\mathcal{T}_r(\mathcal{V}, \mathcal{E}_r)$, i.e., $j_1 < j_2$ if j_1 is on the path from r to j_2.

Consider a cloud of N particles and assume that each particle evolves according to a Markov process. Denote the states of the Markov process by $X = \{X_1, X_2, \dots, X_n\}$ and let the transition probability matrix on edge $(j_1, j_2) \in \mathcal{E}_r$ be $A^{(j_1,j_2)} \in \mathbb{R}_+^{n\times n}$. Let the vector μ_j describe the particle distribution over the discrete state space X on node j, for $j = 1, \dots, J$. In analogy to the optimal transport framework we define the mass transport matrix $M^{(j_1,j_2)}$, where element $M_{k\ell}^{(j_1,j_2)}$ describes the number of particles transitioning from state k to state ℓ on edge (j_1, j_2). Given μ_{j_1} and $A_{(j_1,j_2)}$, the probability for the event that a given mass transfer matrix $M^{(j_1,j_2)}$ describes the underlying particle dynamics satisfies a large deviation principle with rate function $H(\cdot \,|\, \mathrm{diag}(\mu_{j_1})A^{(j_1,j_2)})$, i..e.,

$$\mathbb{P}_{\mu_{j_1},A^{(j_1,j_2)}}(M^{(j_1,j_2)}) \sim e^{-H\left(M^{(j_1,j_2)}|\mathrm{diag}(\mu_j)A^{(j_1,j_2)}\right)},$$

where $H(P \,|\, Q) := \sum_{i,j} \left(P_{i,j} \log\left(P_{ij}/Q_{ij}\right) - P_{ij} + Q_{ij}\right)$ is the normalized KL divergence between two matrices P and Q. Thus, the discrete Schrödinger bridge problem in Pavon and Ticozzi (2010) (see also Haasler et al. (2019)) can be naturally extended to the tree structure as

$$\begin{aligned} \underset{\substack{M^{(j_1,j_2)},\,(j_1,j_2)\in\mathcal{E}_r,\\ \mu_j,\,j\in\mathcal{V}\setminus\Gamma}}{\text{minimize}} \quad & \sum_{(j_1,j_2)\in\mathcal{E}_r} H\left(M^{(j_1,j_2)} \,|\, \mathrm{diag}(\mu_{j_1})A^{(j_1,j_2)}\right) \\ \text{subject to} \quad & M^{(j_1,j_2)}\mathbf{1} = \mu_{j_1},\ (M^{(j_1,j_2)})^T\mathbf{1} = \mu_{j_2}, \\ & \text{for} \ \ (j_1,j_2) \in \mathcal{E}_r. \end{aligned} \tag{6}$$

It can be shown that under certain conditions the solution to the generalized Schrödinger bridge problem (6) is equivalent to the solution of the entropy regularized multi-marginal optimal transport problem (1) on the same tree. In particular, if the cost matrices in (4) and the transition probabilities for problem (6) are chosen such that

$$C^{(j_1,j_2)} = -\epsilon \log(A^{(j_1,j_2)}), \ \text{for all } (j_1,j_2) \in \mathcal{E}_r, \tag{7}$$

then it holds for the optimizers of (1) and (6) that

$$P_{(j_1,j_2)}(\mathbf{M}) = M^{(j_1,j_2)}, \qquad \text{for } (j_1,j_2) \in \mathcal{E}_r,$$

where the pairwise projections are defined as

$$P_{j_1,j_2}(\mathbf{M}) = \sum_{i_1,\dots,i_J\setminus\{i_{j_1},i_{j_2}\}} \mathbf{M}_{i_1,\dots,i_J},$$

and consequently that $P_j(\mathbf{M}) = \mu_j$ for all $j \in \mathcal{V}$.

4. PAIRWISE REGULARIZED OPTIMAL TRANSPORT ON A TREE

Another natural way to define an optimal transport problem on the tree $\mathcal{T}$ is to minimize the sum of all bi-marginal transport costs on the edges of $\mathcal{T}$. In fact, barycenter problems are typically formulated in this pairwise manner, see, e.g., Cuturi and Doucet (2014); Elvander et al. (2019). Although, in the bi-marginal case the entropy regularized optimal transport problem is equivalent to the Schrödinger bridge, we show that this equivalence does not extend to the respective problems defined on trees.

Given the cost matrices $C^{(j_1,j_2)} \in \mathbb{R}_+^{n\times n}$ for $(j_1, j_2) \in \mathcal{E}$, the pairwise entropy regularized optimal transport problem on $\mathcal{T}$ is defined as

$$\underset{\mu_j,\,j\in\mathcal{V}\setminus\Gamma}{\text{minimize}} \quad \sum_{(j_1,j_2)\in\mathcal{E}} T_\epsilon^{(j_1,j_2)}(\mu_{j_1},\mu_{j_2}), \tag{8}$$

where

$$\begin{aligned} T_\epsilon(\mu_1,\mu_2) = \underset{M\in\mathbb{R}_+^{n\times n}}{\text{minimize}} \quad & \mathrm{trace}((C^{(j_1,j_2)})^T M) + \epsilon H(M) \\ \text{subject to} \quad & M\mathbf{1} = \mu_1, \quad M^T\mathbf{1} = \mu_2. \end{aligned}$$

We note that the objective function of the generalized Schrödinger bridge (6) can be written as

$$\sum_{(j_1,j_2)\in\mathcal{E}_r} H\left(M^{(j_1,j_2)} \,|\, A^{(j_1,j_2)}\right) - \sum_{j\in\mathcal{V}\setminus\mathcal{L}} (\deg(j)-1)H(\mu_j).$$

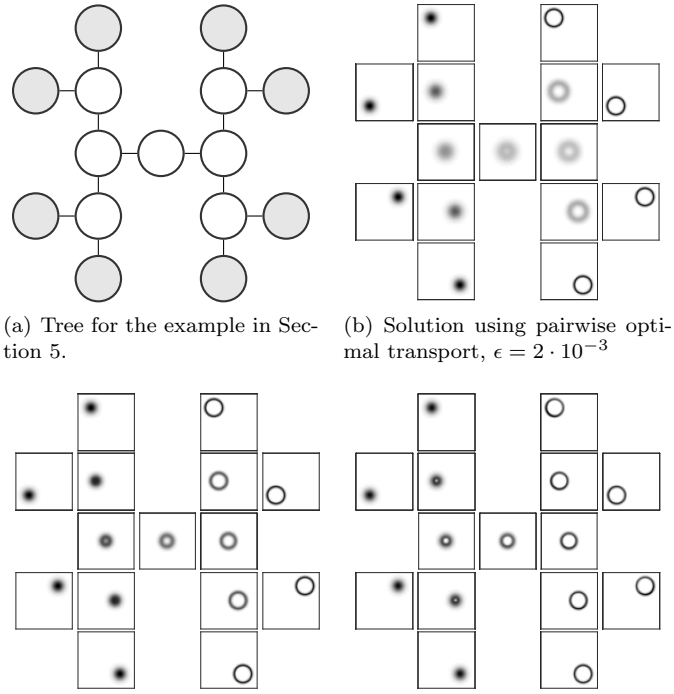
If relation (7) holds, the generalized Schrödinger bridge and entropy regularized multi-marginal optimal transport problem are equivalent, and can be written as

$$\underset{\mu_j,\,j\in\mathcal{V}\setminus\Gamma}{\text{minimize}} \sum_{(j_1,j_2)\in\mathcal{E}} T_\epsilon^{(j_1,j_2)}(\mu_{j_1},\mu_{j_2}) - \sum_{j\in\mathcal{V}\setminus\mathcal{L}} (\deg(j)-1)H(\mu_j).$$

Thus, the multi-marginal optimal transport problem penalizes not only the transport cost between the marginals, but in addition favors marginal distributions with high entropy. One can thus expect less smoothed out distributions when solving the multi-marginal optimal transport problem, which is desirable in many applications, such as the localization problems in Elvander et al. (2020) and computer vision applications (e.g., the ones in Solomon et al. (2015)). Note that this qualitative difference between the pairwise and multi-marginal formulation has been previously observed, yet not fully explained, for a barycenter optimal transport problem, i.e., a star graph in Elvander et al. (2020). Moreover, empirical study suggests that the multi-marginal problem is better conditioned compared to the pairwise problem, which allows for smaller values of the regularization parameter ϵ, while still yielding a numerically stable algorithm.

5. EXAMPLE

In this section we compare the solutions to the entropy regularized multi-marginal and pairwise optimal transport problems on the tree $\mathcal{T}$ illustrated in Figure 1(a). The marginals on the 15 nodes are all 50×50 pixel image, and the marginal images on the 8 leaves, colored in gray in Figure 1(a), are known. Each edge on the tree is associated with a cost function defined by the L_2-distance between any two pixels. Using this choice of cost function in the



(a) Tree for the example in Section 5.

(b) Solution using pairwise optimal transport, $\epsilon = 2 \cdot 10^{-3}$

(c) Solution using multi-marginal optimal transport, $\epsilon = 2 \cdot 10^{-3}$

(d) Solution using multi-marginal optimal transport, $\epsilon = 4 \cdot 10^{-4}$

Fig. 1. Estimated marginals of the pairwise (b) and multi-marginal (c,d) optimal transport solutions on the tree in (a).

corresponding optimal transport problems yields smooth translations of power for the intermediate marginals.

We solve the pairwise entropy regularized optimal transport problem (8) with regularization parameter $\epsilon = 2 \cdot 10^{-3}$ on $\mathcal{T}$. The solution can be seen in Figure 1(b). Compared to the pairwise optimal transport estimate, the solution to the entropy regularized multi-marginal optimal transport problem (1) on the same tree $\mathcal{T}$ and with the same regularization parameter ϵ is significantly sharper and less smoothed out, see Figure 1(c). For the pairwise optimal transport problem, the method diverges with a smaller regularization parameter, e.g., $\epsilon = 10^{-3}$. In contrast, for the multi-marginal formulation the regularization parameter can be decreased further, still yielding a numerically stable algorithm. We have found that the method is still stable for a regularization parameter of $\epsilon = 4 \cdot 10^{-4}$, which results in very clear estimates on the intermediate nodes.

REFERENCES

Adler, J., Ringh, A., Öktem, O., and Karlsson, J. (2017). Learning to solve inverse problems using Wasserstein loss. *arXiv preprint arXiv:1710.10898*.

Benamou, J.D., Carlier, G., Cuturi, M., Nenna, L., and Peyr, G. (2015). Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37(2), A1111A1138.

Chen, Y., Georgiou, T., and Pavon, M. (2016a). Entropic and Displacement Interpolation: A Computational Approach Using the Hilbert Metric. *SIAM Journal on Applied Mathematics*, 76(6), 2375–2396.

Chen, Y., Georgiou, T., and Pavon, M. (2016b). On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2), 671–691.

Chen, Y., Georgiou, T., and Pavon, M. (2016c). Optimal Steering of a Linear Stochastic System to a Final Probability Distribution, Part I. *IEEE Transactions on Automatic Control*, 61(5), 1158–1169.

Chen, Y. and Karlsson, J. (2018). State tracking of linear ensembles via optimal mass transport. *IEEE Control Systems Letters*, 2(2), 260–265.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, 2292–2300. Neural Information Processing Systems Foundation.

Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, 685–693.

Dominitz, A. and Tannenbaum, A. (2010). Texture Mapping via Optimal Mass Transport. *IEEE Transactions on Visualization and Computer Graphics*, 16(3), 419–433.

Elvander, F., Haasler, I., Jakobsson, A., and Karlsson, J. (2019). Non-Coherent Sensor Fusion via Entropy Regularized Optimal Mass Transport. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4415–4419.

Elvander, F., Haasler, I., Jakobsson, A., and Karlsson, J. (2020). Multi-Marginal Optimal Mass Transport using Partial Information with Applications in Robust Localization and Sensor Fusion. *Signal Processing*.

Georgiou, T. and Pavon, M. (2015). Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics*, 56(3), 033301.

Haasler, I., Ringh, A., Chen, Y., and Karlsson, J. (2019). Estimating ensemble flows on a hidden Markov chain. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 1331–1338. IEEE.

Haasler, I., Ringh, A., Chen, Y., and Karlsson, J. (2021). Multimarginal optimal transport with a tree-structured cost and the Schrödinger bridge problem. *SIAM Journal on Control and Optimization*, 59(4), 2428–2453.

Léonard, C. (2014). A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems - A*, 34(4), 1533–1574.

Pass, B. (2015). Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6), 1771–1790.

Pavon, M. and Ticozzi, F. (2010). Discrete-time classical and quantum Markovian evolutions: Maximum entropy problems on path space. *Journal of Mathematical Physics*, 51(4), 042104.

Schrödinger, E. (1931). Über die Umkehrung der Naturgesetze. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse*, 144–153.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4), 66:1–66:11.

Villani, C. (2008). *Optimal transport: Old and new*. Springer, Berlin Heidelberg.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Splitting algorithms and circuits analysis

Thomas Chaffey * **Amritam Das** ** **Rodolphe Sepulchre** *

* *University of Cambridge, Department of Engineering, Trumpington Street, Cambridge CB2 1PZ,* {tlc37, rs771}@cam.ac.uk.
** *Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden,* amritam@kth.se.

Abstract: The splitting algorithms of monotone operator theory find zeros of sums of relations. This corresponds to solving series or parallel one-port electrical circuits, or the negative feedback interconnection of two subsystems. One-port circuits with series *and* parallel interconnections, or block diagrams with multiple forward and return paths, give rise to current-voltage relations consisting of nested sums *and inverses*. In this extended abstract, we present new splitting algorithms specially suited to these structures, for interconnections of monotone and anti-monotone relations.

Keywords: Scaled Relative Graph, Nyquist, loop shaping, robustness

1. INTRODUCTION

The mathematical property of *monotonicity* originated in the study of networks of nonlinear resistors (Duffin, 1946; Zarantonello, 1960; Dolph, 1961; Minty, 1960, 1961a,b). Monotonicity generalizes the concept of passivity from linear circuit theory; loosely speaking, an element is monotone if it is passive with respect to *any* possible reference trajectory. Following the influential paper of Rockafellar (1976), monotone operator theory has grown to become a pillar of large scale optimization theory (Bauschke and Combettes, 2011; Ryu and Yin, 2022; Parikh and Boyd, 2013; Bertsekas, 2011).

Central to this theory are the family of *splitting algorithms*. These algorithms find zeros of sums of monotone operators, and allow computation to be performed separately for each operator. Recent work by the authors has revisited the study of electrical networks using modern splitting algorithms (Chaffey and Sepulchre, 2021). The main idea is that finding a zero of the sum of two operators is equivalent to solving the port behavior of their parallel (or series) interconnection. In turn, this is equivalent to solving the behavior of the negative feedback interconnection of two elements. This observation motivates the development of splitting algorithms which match more general circuit architectures. In Section 4, we describe an algorithm which solves the behavior of arbitrary series/parallel one-port circuits.

While splitting methods require each circuit element to be monotone, similar ideas can be applied to *mixed monotone* circuits, consisting of port interconnections of monotone and anti-monotone elements. This significantly expands the possible types of circuit behavior, allowing, for example, relaxation oscillations (van der Pol, 1926) and neuronal excitability (FitzHugh, 1961). In (Das et al., 2021), the authors have adapted Difference of Convex Programming (Lipp and Boyd, 2016; Yuille and Rangarajan, 2003) to solve such behaviors. In Section 5, we describe a new splitting algorithm which matches the mixed feedback structure of oscillators such as the van der Pol and FitzHugh-Nagumo models.

While classical splitting methods deal only with sums, the algorithms we describe here deal with both sums and inverses - the two operations which constitute physical port interconnections. The algorithms described in this abstract form the basis for a more general class of splitting algorithms, which correspond to arbitrary interconnections of physical systems.

2. MONOTONE AND ANTI-MONOTONE RELATIONS

A Hilbert space $\mathcal{H}$ is a complete vector space equipped with an inner product, $\langle\cdot|\cdot\rangle : \mathcal{H}\times\mathcal{H}\to\mathbb{C}$, and an induced norm $\|x\| \coloneqq \sqrt{\langle x|x\rangle}$. In this abstract, we will treat general Hilbert spaces, although a common choice in practice is the space of square-summable, discrete time signals on $[0,T]$, denoted $l_{2,T}$.

An *operator* on $\mathcal{H}$, is a possibly multi-valued map $R:\mathcal{H}\to\mathcal{H}$. The identity operator, which maps $u\in\mathcal{X}$ to itself, is denoted by I. The domain of an operator R is denoted $\operatorname{dom} R$. The *graph*, or *relation*, of an operator, is the set $\{u,y \mid u\in\operatorname{dom} R, y\in R(u)\}\subseteq\mathcal{H}\times\mathcal{H}$. We use the notions of an operator and its relation interchangeably, and denote them in the same way.

The standard operations on functions can be extended to relations. Let R and S be relations on an arbitrary Hilbert space $\mathcal{H}$. Then:

$$S^{-1} = \{(y,u) \mid y\in S(u)\}$$
$$S+R = \{(x,y+z) \mid (x,y)\in S, (x,z)\in R\}$$
$$SR = \{(x,z) \mid \exists\, y \text{ s.t. } (x,y)\in R, (y,z)\in S\}.$$

* The research leading to these results has received funding from the European Research Council under the Advanced ERC Grant Agreement Switchlet n. 670645, and from the Cambridge Philosophical Society.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Note that S^{-1} always exists, but is not an inverse in the usual sense. In particular, in general $S^{-1}S \neq I$.

Definition 1. A relation $S \subseteq \mathcal{H} \times \mathcal{H}$ is called *monotone* if

$$\langle u_1 - u_2 | y_1 - y_2 \rangle \geq 0$$

for any $(u_1, y_1), (u_2, y_2) \in S$. A monotone relation is called *maximal* if it is not properly contained in any other monotone relation.

Definition 2. A relation $S : \mathcal{H} \to \mathcal{H}$ is *anti-monotone* if $-S$ is monotone.

3. SPLITTING TWO-ELEMENT CIRCUITS

There is a large body of literature on splitting algorithms, which solve problems of the form $0 \in M_1(u) + M_2(u)$, where M_1 and M_2 are maximal monotone relations. There is a direct analogy with electrical circuits: if M_1 and M_2 are resistances, their series interconnection is given by the relation $v = M_1(i) + M_2(i)$; if M_1 and M_2 are conductances, their parallel interconnection is given by $i = M_1(v) + M_2(v)$. Given a current, the corresponding voltage across a parallel interconnection can be found using a splitting algorithm, by solving $0 \in M_1(v) + M_2(v) - i$. Here, we briefly describe two splitting algorithms – the forward/backward splitting, and the Douglas-Rachford splitting. For the convergence properties of these algorithms, we refer the reader to (Giselsson and Moursi, 2019; Bauschke and Combettes, 2011; Ryu and Yin, 2022). Given an operator S and a scaling factor α, the α-resolvent of S is defined to be the operator

$$\operatorname{res}_{\alpha S} \coloneqq (I + \alpha S)^{-1}.$$

If S is maximal monotone, res_S is single-valued (Minty, 1961a).

3.1 Forward/backward splitting

The simplest splitting algorithm is the forward/backward splitting (Passty, 1979; Gabay, 1983; Tseng, 1988). Suppose M_1 and $\operatorname{res}_{\alpha M_2}$ are single-valued. Then:

$$\begin{aligned} & 0 \in M_1(x) + M_2(x) \\ \iff \quad & 0 \in x - \alpha M_1(x) - (x + \alpha M_2(x)) \\ \iff \quad & (I + \alpha M_2)x \ni (I - \alpha M_1)x \\ \iff \quad & x = \operatorname{res}_{\alpha M_2}(I - \alpha M_1)x. \end{aligned}$$

The fixed point iteration $x^{j+1} = \operatorname{res}_{\alpha M_2}(x^j - \alpha M_1(x^j))$ is the forward/backward splitting algorithm.

3.2 Douglas-Rachford splitting

The reflected resolvent, or Cayley operator, is the operator

$$R_{\alpha S} \coloneqq 2\operatorname{res}_{\alpha S} - I.$$

Given two operators M_1 and M_2, and a scaling factor α, the Douglas-Rachford algorithm (Douglas and Rachford, 1956; Lions and Mercier, 1979) is the iteration

$$\begin{aligned} z^{k+1} &= T(z^k), \\ x^k &= \operatorname{res}_{\alpha M_2}(z^k), \end{aligned}$$

where T is given by

$$T = \frac{1}{2}(I + R_{\alpha M_1} R_{\alpha M_2}). \tag{1}$$

4. SPLITTING N-ELEMENT CIRCUITS

If our circuit is composed of three elements, with one series interconnection and one parallel interconnection (Figure 1), it has the form $M = M_1 + (M_2 + M_3)^{-1}$. A naive approach to solving the behavior of this circuit is to use a splitting algorithm such as the forward/backward algorithm, with the resolvent step applied for M_1 and the forward step applied for $(M_2 + M_3)^{-1}$. Applying this forward step amounts to solving $v = (M_2 + M_3)^{-1}(i)$ for some u, which may be rewritten as $0 \in (M_2 + M_3)(v) - i$. This can be solved by again applying the forward/backward algorithm.

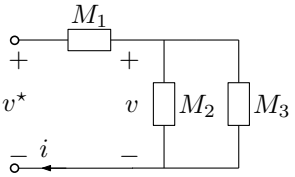


Fig. 1. Three elements with one series interconnection and one parallel interconnection.

This naive procedure has poor complexity: for every forward/backward step for $M_1 + (M_2 + M_3)^{-1}$, an *entire* fixed point iteration has to be computed for (an offset version of) $M_2 + M_3$. In (Chaffey and Sepulchre, 2021), we propose an alternative procedure for n-element circuits. Here, we sketch this procedure on the circuit of Figure 1. Rather than apply a forward step for the relation $(M_2 + M_3)^{-1}$, we simply apply a *single step* of the fixed point iteration needed to compute this forward step, using the forward/backward algorithm. Given $v^\star$, we want to solve $0 \in (M_1 + (M_2 + M_3)^{-1})(i) - v^\star$. Assume that M_3, $\operatorname{res}_{\alpha_1 M_2}$ and $\operatorname{res}_{\alpha_2 M_1}$ are single-valued. We then have:

$$v^\star \in v + M_1(i) \tag{2}$$

$$v \in (M_2 + M_3)^{-1}(i), \tag{3}$$

where v is the voltage over M_2, illustrated in Figure 1. Equation (2) gives

$$\begin{aligned} i + \alpha_2 M_1(i) &\ni i - \alpha_2 v + \alpha_2 v^\star \\ i &= (I + \alpha_2 M_1)^{-1}(i - \alpha_2 v + \alpha_2 v^\star) \\ i &= \operatorname{res}_{\alpha_2 M_1}(i - \alpha_2 v + \alpha_2 v^\star). \end{aligned}$$

Equation (3) gives

$$\begin{aligned} i &\in (M_2 + M_3)(v) \\ v + \alpha_1 M_2(v) &\ni v - \alpha_1 M_3(v) + \alpha_1 i \\ v &= (I + \alpha_1 M_2)^{-1}(v - \alpha_1 M_3(v) + \alpha_1 i) \\ v &= \operatorname{res}_{\alpha_1 M_2}(v - \alpha_1 M_3(v) + \alpha_1 i). \end{aligned}$$

This shows that a fixed point of the iteration

$$\begin{aligned} v^{k+1} &= \operatorname{res}_{\alpha_1 M_2}(v^k - \alpha_1 M_3(v^k) + \alpha_1 i^k) \\ i^{k+1} &= \operatorname{res}_{\alpha_2 M_1}(i^k - \alpha_2 v^{k+1} + \alpha_2 v^\star) \end{aligned}$$

is a solution to our original problem $0 \in (M_1 + (M_2 + M_3)^{-1})(i) - v^\star$.

5. SPLITTING THE DIFFERENCE

A mixture of positive and negative feedback is a ubiquitous mechanism, in both biology and engineering, for the generation of switches and oscillations (Sepulchre et al., 2019; Sepulchre and Stan, 2005; Stan et al., 2007; Stan and Sepulchre, 2007; Chua et al., 1987). Again adopting the

analogy of electrical circuits, such feedback systems can be thought of as the parallel interconnection of three elements (Figure 2) - the forward path and the negative feedback path, which we assume to be monotone, and the positive feedback path, which we assume to be anti-monotone. Such a structure encompasses systems such as the van der Pol and FitzHugh-Nagumo oscillators. For example, the van der Pol oscillator is given by $A_1(s) = (s^2+1)/s$, $A_2(v) = \mu v^3/3$ and $B(v) = \mu v$ (where s is the Laplace variable) (Das et al., 2021).

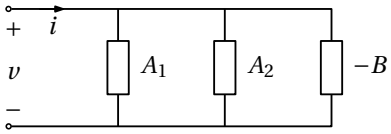


Fig. 2. A parallel mixed monotone circuit, which is a prototype structure for systems such as the van der Pol and FitzHugh-Nagumo oscillators.

Given the mixed monotone structure of Figure 2, we can find the steady state behavior of the system by solving a zero-finding problem: $0 \in A_1(v) + A_2(v) - B(v) - i$.

The authors have explored methods to solve these problems using an adaptation of Difference of Convex Programming in (Das et al., 2021). The method involves iterating the operator $(A_1 + A_2)^{-1}B$. Computing $(A_1 + A_2)^{-1}$ at every iteration is an expensive operation; in this section, we propose the *mixed monotone Douglas-Rachford algorithm* (Algorithm 7), which replaces $(A_1 + A_2)^{-1}$ with a single step of the Douglas-Rachford iteration needed to invert it.

For operators A_1 and A_2 and step size α, we define $T_\alpha(A_1, A_2)$ to be the Douglas-Rachford operator:

$$T_\alpha(A_1, A_2) = \frac{1}{2}(I + R_{\alpha A_1} R_{\alpha A_2}). \tag{4}$$

Recall that $R_{\alpha S}$ denotes the Cayley operator $2\mathrm{res}_{\alpha S} - I$.

Algorithm 1 Mixed-Monotone Douglas-Rachford

1: **Data:** Maximal monotone A_1, A_2. Monotone, single-valued B. Initial value x_1. Convergence tolerance $\varepsilon > 0$.
2: Define A_1^j by $x \mapsto A_1(x) - y_j$ for all j.
3: $j = 1$
4: **do**
5: Solve

$$x_{j+1} = \mathrm{res}_{\alpha A_2}(z_j)$$
$$y_{j+1} = B(x_{j+1})$$
$$z_{j+1} = T_\alpha(A_1^{j+1}, A_2 a)(z_j).$$

6: $j = j + 1$.
7: **while** $|x_{j+1} - x_j| > \varepsilon$

Note that a fixed point of this algorithm is a solution to $0 \in A_1(x) + A_2(x) - B(x)$: we know, by convergence of the Douglas-Rachford algorithm, that x is a solution to $0 \in A_1^j(x) + A_2(x)$, which is equal to $A_1(x) + A_2(x) - B(x)$ at a fixed point. (Chaffey, 2022, Thm. 4.1) gives a convergence condition for this algorithm. Figure 3 shows steady-state solutions to the van der Pol oscillator computed with Algorithm 7. For further details of the implementation, the reader is referred to (Chaffey, 2022, Example 4.3).

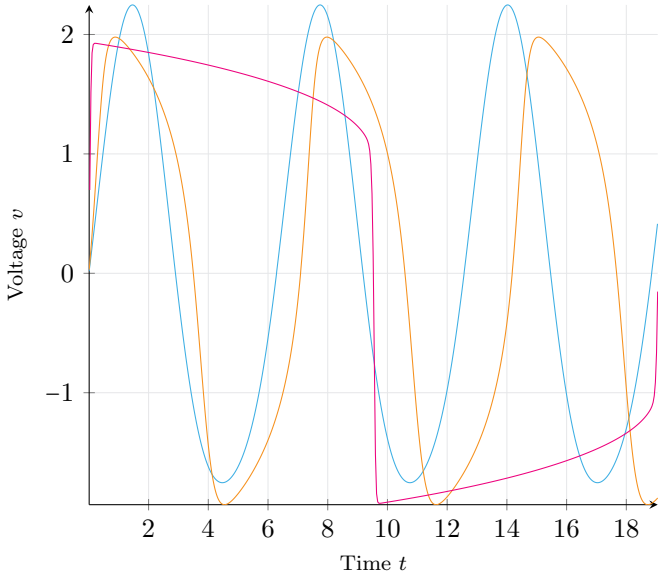


Fig. 3. Steady-state solutions to the van der Pol oscillator for $\mu = 0.0002$ (blue), 1.5 (orange) and 10 (red). Algorithmic parameters are a step size of $\alpha = 0.05$, convergence tolerance of $\varepsilon = 0.01$ and 5000 time steps.

REFERENCES

Bauschke, H.H. and Combettes, P.L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, New York, NY. doi:10.1007/978-1-4419-9467-7.

Bertsekas, D.P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2), 163–195. doi:10.1007/s10107-011-0472-0.

Chaffey, T. (2022). *Input/Output Analysis: Graphical and Algorithmic Methods*. Ph.D. thesis, University of Cambridge.

Chaffey, T. and Sepulchre, R. (2021). Monotone Circuits. In *Proceedings of the European Control Conference*.

Chua, L.O., Desoer, C.A., and Kuh, E.S. (1987). *Linear and Nonlinear Circuits*. McGraw-Hill Series in Electrical Engineering. McGraw-Hill, New York.

Das, A., Chaffey, T., and Sepulchre, R. (2021). Oscillations in Mixed-Feedback Systems. *arXiv:2103.16379 [cs, eess]*.

Dolph, C.L. (1961). Recent developments in some non-self-adjoint problems of mathematical physics. *Bulletin of the American Mathematical Society*, 67(1), 1–70. doi: 10.1090/S0002-9904-1961-10493-X.

Douglas, J. and Rachford, H.H. (1956). On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables. *Transactions of the American Mathematical Society*, 82(2), 421–439. doi: 10.2307/1993056.

Duffin, R.J. (1946). Nonlinear networks. I. *Bulletin of the American Mathematical Society*, 52(10), 833–839. doi: 10.1090/S0002-9904-1946-08650-4.

FitzHugh, R. (1961). Impulses and Physiological States in Theoretical Models of Nerve Membrane. *Biophysical Journal*, 1(6), 445–466. doi:10.1016/s0006-3495(61)86902-6.

Gabay, D. (1983). Chapter IX Applications of the Method of Multipliers to Variational Inequalities. In M. Fortin and R. Glowinski (eds.), *Studies in Mathematics and*

Its Applications, volume 15 of *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. Elsevier. doi:10.1016/S0168-2024(08)70034-1.

Giselsson, P. and Moursi, W.M. (2019). On compositions of special cases of Lipschitz continuous operators. *arXiv:1912.13165 [math]*.

Lions, P.L. and Mercier, B. (1979). Splitting Algorithms for the Sum of Two Nonlinear Operators. *SIAM Journal on Numerical Analysis*, 16(6), 964–979. doi:10.1137/0716071.

Lipp, T. and Boyd, S. (2016). Variations and extension of the Convex–Concave procedure. *Optimization and Engineering*, 17(2), 263–287. doi:10.1007/s11081-015-9294-x.

Minty, G.J. (1960). Monotone networks. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 257(1289), 194–212. doi:10.1098/rspa.1960.0144.

Minty, G.J. (1961a). On the maximal domain of a "monotone" function. *The Michigan Mathematical Journal*, 8(2), 135–137. doi:10.1307/mmj/1028998564.

Minty, G.J. (1961b). Solving Steady-State Nonlinear Networks of 'Monotone' Elements. *IRE Transactions on Circuit Theory*, 8(2), 99–104. doi:10.1109/TCT.1961.1086765.

Parikh, N. and Boyd, S. (2013). Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3), 123–231.

Passty, G.B. (1979). Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2), 383–390. doi:10.1016/0022-247X(79)90234-8.

Rockafellar, R.T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5), 877–898. doi:10.1137/0314056.

Ryu, E.K. and Yin, W. (2022). *Large-Scale Convex Optimization via Monotone Operators*. Draft edition.

Sepulchre, R., Drion, G., and Franci, A. (2019). Control Across Scales by Positive and Negative Feedback. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1), 89–113. doi:10.1146/annurev-control-053018-023708.

Sepulchre, R. and Stan, G.B. (2005). Feedback mechanisms for global oscillations in Lure systems. *Systems & Control Letters*, 54(8), 809–818. doi:10.1016/j.sysconle.2004.12.004.

Stan, G.B., Hamadeh, A., Sepulchre, R., and Goncalves, J. (2007). Output synchronization in networks of cyclic biochemical oscillators. In *IEEE American Control Conference*, 3973–3978. doi:10.1109/ACC.2007.4282673.

Stan, G.B. and Sepulchre, R. (2007). Analysis of Interconnected Oscillators by Dissipativity Theory. *IEEE Transactions on Automatic Control*, 52(2), 256–270. doi:10.1109/tac.2006.890471.

Tseng, P. (1988). Applications of a Splitting Algorithm to Decomposition in Convex Programming and Variational Inequalities. *SIAM Journal on Control and Optimization*, 29(1), 119–138. doi:10.1137/0329006.

van der Pol, B. (1926). On "Relaxation-Oscillations". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 978–992. doi:10.1080/14786442608564127.

Yuille, A.L. and Rangarajan, A. (2003). The Concave-Convex Procedure. *Neural Computation*, 15(4), 915–936. doi:10.1162/08997660360581958.

Zarantonello, E.H. (1960). Solving Functional Equations by Contractive Averaging. Technical Report PB166988, Mathematics Research Center, Univ. of Wisconsin, Madison.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Combining the SOS and SONC cones - A Hilbert's 1888 Theorem analogue and further separation results

Moritz Schick *

* *University of Konstanz, Konstanz, Germany (e-mail: moritz.schick@uni-konstanz.de).*

Abstract: Studying convex cones inside the cone of positive semidefinite (PSD) polynomials is an important field of research in real algebraic geometry and polynomial optimization. In this work, we combine two such well established cones, which are sums of squares (SOS) and sums of nonnegative circuit polynomials (SONC) and consider PSD polynomials, that decompose into an SOS and a SONC part. We call the resulting set the SOS+SONC cone. For this newly established cone, we prove two separation results. The first one is an analogue to Hilbert's 1888 Theorem for the SOS+SONC cone. The second one shows that whenever the SOS and SONC cones are proper subsets of the PSD cone, they are also proper subsets of the SOS+SONC cone.

Keywords: sums of squares, sums of nonnegative circuit polynomials, Hilbert's 1888 Theorem, polynomial optimization

1. INTRODUCTION

Minimizing a given real, multivariate polynomial $f \in \mathbb{R}[\mathbf{x}]$ is the central challenge of polynomial optimization. The importance of this topic can be seen in the variety of applications in different fields such as optimal control, mathematical finance and real-time decision making. It is well known that polynomial optimization can equivalently be viewed as the problem of deciding nonnegativity of real polynomials. This equivalence is of central meaning in real algebraic geometry since convex geometric tools can be used to obtain a deeper understanding of the set of nonnegative polynomials.

The relevance of a theoretical study of both problems is also stressed by the fact that both polynomial optimization and deciding nonnegativity of real polynomials are in general NP-hard even for low dimensional cases. Hence, one is often interested in solving easier problems instead, often involving trade offs between feasibility and preciseness of solutions. In this work we follow the real algebraic geometric approach of taking suitable inner approximations of the set of nonnegative polynomials.

A first inner approximation of the cone of nonnegative polynomials are *sums of squares (SOS)*, which have a long history in Mathematics and go back to Hilbert's seminal work in Hilbert (1888). The SOS approach has proven to be a powerful tool for solving a vast number of optimization problems, see see e.g. Lasserre (2009) for more details. However, it has its limitations especially in high degree and high number of variables cases.

A second approximation which has gained a lot of interest in recent years are *sums of nonnegative circuit (SONC) polynomials*, which were first introduced by Iliman and de Wolff (2016). The sparse structure of this class of polynomials allows to solve large problems, where the SOS approach has its difficulties. However, since the SONC approach is relatively new, it is only fully developed for special classes of polynomials, having e.g. simplex Newton polytopes. Indeed, there are different types of conic programming using SONC for polynomial optimization, see e.g. Wang and Magron (2020) and Dressler et al. (2020).

Dressler (2018) pointed out that if it would be possible to combine the two approaches and use the best of both worlds, one would get a new approximation which is at least as good as the single approaches themselves.

In terms of polynomial optimization, let $f \in \mathbb{R}[\mathbf{x}]$ and consider the global polynomial optimization problem $f^* = \inf_{x\in\mathbb{R}^n} f(x) = \sup\{\lambda \in \mathbb{R} : f - \lambda \geq 0 \text{ on } \mathbb{R}^n\}$. Let SOS and SONC denote the sets of SOS and SONC polynomials, respectively. Then, lower bounds on f^* can be achieved via $f_{\text{SOS}} := \sup\{\lambda \in \mathbb{R} : f - \lambda \in \text{SOS}\}$ and $f_{\text{SONC}} := \sup\{\lambda \in \mathbb{R} : f - \lambda \in \text{SONC}\}$. Taking the Minkowski sum $\text{SOS} + \text{SONC}$ leads to a third lower bound $f_{\text{SOS+SONC}} := f_{\text{SOS}+\text{SONC}} := \sup\{\lambda \in \mathbb{R} : f - \lambda \in \text{SOS} + \text{SONC}\}$ which satisfies $f_{\text{SOS}}, f_{\text{SONC}} \leq f_{\text{SOS+SONC}} \leq f^*$.

In this work, we fully characterize the numbers of variables and degrees of polynomials for which the above's inequalities are strict. Therefore, we formally introduce the cone of sums of squares and nonnegative polynomials (SOS+SONC) and present explicit examples of polynomials separating this cone from the SOS and SONC cones as well as the cone of positive semidefinite polynomials.

2. PRELIMINARIES

2.1 Notations

Let $\mathbb{N} := \{1, 2, 3, \ldots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ be the sets of positive and nonnegative integers, respectively and $[m] := \{1, \ldots, m\}$ $(m \in \mathbb{N})$. For $n \in \mathbb{N}$ let $\mathbb{R}[\mathbf{x}] := \mathbb{R}[\mathrm{x}_1, \ldots, \mathrm{x}_n]$

be the polynomial ring over $\mathbb{R}$ in n variables. The integer $n \in \mathbb{N}$ will be fixed throughout this work.

A polynomial $f \in \mathbb{R}[\mathbf{x}]$ with all monomials having the same degree $k \in \mathbb{N}$ is called *homogeneous* or a *form*. For $k \in \mathbb{N}$ we denote by $H_{n,k}$ the finite dimensional vector space of n-variate, homogeneous polynomials of degree k. For $f \in \mathbb{R}[\mathbf{x}]$, we write $\operatorname{supp}(f)$ for the *support* and $\operatorname{New}(f)$ for the *Newton polytope* of f, i.e. $\operatorname{New}(f) = \operatorname{conv}(\operatorname{supp}(f))$, where $\operatorname{conv}(S)$ is the *convex hull* of a set S. For an arbitrary polytope $\Delta \subseteq \mathbb{R}^n$, we denote its set of *vertices* by $V(\Delta)$. If $\Delta = \operatorname{New}(f)$ is the Newton polytope of a polynomial f, we also write $V(f) := V(\operatorname{New}(f))$. For $\alpha \in \mathbb{N}_0^n$ and $f \in \mathbb{R}[\mathbf{x}]$, we denote by f_α the coefficient of f corresponding to x^α, i.e. $f = \sum_{\alpha \in \mathbb{N}_0^n} f_\alpha \mathrm{x}^\alpha$ and $f_\alpha = 0$ if $\alpha \notin \operatorname{supp}(f)$.

2.2 Young's Inequality

The following Theorem is essential for proofs in Section 3.

Young's Inequality Let $1 < p, q < \infty$ be s.t. $1/p + 1/q = 1$. Further, let $a, b \in \mathbb{R}$ be arbitrary. Then $ab \leq \frac{|a|^p}{p} + \frac{|b|^q}{q}$. Further, for $a, b \geq 0$ equality holds if and only if $a^p = b^q$.

2.3 Positive Semidefinite (PSD) Polynomials

A polynomial $f \in \mathbb{R}[\mathbf{x}]$ is *positive semidefinite (PSD)* if $f(x) \geq 0$ for all $x \in \mathbb{R}^n$. Clearly, a PSD polynomial must have even degree. Therefore, an even integer $2d \in 2\mathbb{N}$ will be fixed throughout this work.

We denote by

$$P_{n,2d} := \{f \in H_{n,2d} : f \geq 0 \text{ on } \mathbb{R}^n\}$$

the set of PSD forms of degree $2d$. It is well known that $P_{n,2d}$ is a closed, convex cone in the vector space $H_{n,2d}$.

2.4 Sum of Squares (SOS) Polynomials

A polynomial $f \in \mathbb{R}[\mathbf{x}]$ is a *sum of squares (SOS)* if it admits a decomposition of the form $f = \sum_{i=1}^s f_i^2$ such that $s \in \mathbb{N}$, $f_i \in \mathbb{R}[\mathbf{x}]$ $(i \in [s])$. The set of SOS forms of degree $2d$ is denoted by

$$\Sigma_{n,2d} := \left\{ f = \sum_{i=1}^s f_i^2 : f_i \in \mathbb{R}[\mathbf{x}]_{\leq d}, s \in \mathbb{N} \right\}.$$

Similar as the PSD cone, $\Sigma_{n,2d}$ forms a closed, convex cone inside $P_{n,2d}$.

The following theorem characterizes precisely the cases of $(n, 2d)$ for which the PSD and SOS cones coincide. It goes back to Hilbert's work Hilbert (1888).

Hilbert 1888 It holds $\Sigma_{n,2d} = P_{n,2d}$ if and only if $n = 2$ or $2d = 2$ or $(n, 2d) = (3, 4)$.

2.5 Sums of Nonnegative Circuit (SONC) Polynomials

A polynomial $f = \sum_{i=1}^m c_i \underline{\mathrm{x}}^{\alpha(i)} + b\underline{\mathrm{x}}^\beta \in \mathbb{R}[\mathbf{x}]$ where $m \in \mathbb{N}$, $\alpha(1), \ldots, \alpha(m)$, $\beta \in \mathbb{N}_0^n$ and $c_1, \ldots, c_m, b \in \mathbb{R}$ is a *circuit polynomial* if it satisfies the following conditions:

(C1) The lattice points $\alpha(1), \ldots, \alpha(m)$ are even, i.e. $\alpha(1), \ldots, \alpha(m) \in (2\mathbb{N}_0)^n$ and affinely independent.
(C2) The coefficients c_i corresponding to the $\alpha(i)$ are positive, i.e. $c_i > 0$ for $i = 1, \ldots, m$.
(C3) The exponent β lies in the interior of the Newton polytope of f.

The monomials $c_i \mathrm{x}^{\alpha(i)}$ $(i = 1, \ldots, m)$ are called *vertex monomials*. In addition, if $b \neq 0$ the monomial $b\mathrm{x}^\beta$ is called *interior monomial*. The set of circuit polynomials supported on $A \subseteq \mathbb{N}_0^n$ is denoted by $\operatorname{Circ}_A \subseteq \mathbb{R}[\underline{\mathrm{x}}]$.

A polynomial $f \in \mathbb{R}[\mathbf{x}]$ is a *sum of nonnegative circuit polynomials (SONC)* if it admits a decomposition of the form $f = \sum_{i=1}^s f_i$ where $s \in \mathbb{N}$ and every f_i is a nonnegative circuit polynomial. We denote by

$$C_{n,2d} := \left\{ f = \sum_{i=1}^s f_i : f_i \in \operatorname{Circ}_{A_i} \cap P_{n,2d}, A_i \subseteq \mathbb{N}_0^n \right\}$$

the set of SONC polynomials of degree at most $2d$. Again, $C_{n,2d}$ is a closed, convex cone inside $P_{n,2d}$.

The following Theorem shows that $C_{n,2d}$ is an inner approximation of $P_{n,2d}$, which is independent of $\Sigma_{n,2d}$. It is a combination of results in (Iliman and de Wolff, 2016, Prop. 7.2) and (Dressler, 2018, Thm. 3.1.2)

Theorem The SONC cone $C_{n,2d}$ satisfies:

(1.) $C_{n,2d} \subseteq \Sigma_{n,2d}$ if and only if $n = 2$ or $2d = 2$ or $(n, 2d) = (3, 4)$.
(2.) $C_{2,2} = \Sigma_{2,2}$ and $\Sigma_{n,2} \not\subseteq C_{n,2}$ for all $n \geq 3$.
(3.) $\Sigma_{n,2d} \not\subseteq C_{n,2d}$ for all $(n, 2d)$ with $2d \geq 4$.

2.6 The combined cone SOS+SONC of sums of squares and nonnegative circuit polynomials

A polynomial $f \in \mathbb{R}[\mathbf{x}]$ is said to be a *sum of squares and nonnegative circuit polynomials (SOS+SONC)* if it has a decomposition of the form $f = g + h$ for some $f \in \Sigma_{n,2d}, g \in C_{n,2d}$. Further, we denote by

$$(\Sigma + C)_{n,2d} := \Sigma_{n,2d} + C_{n,2d}$$

the set of SOS+SONC forms in n variables of degree $2d$.

3. SEPARATING THE PSD CONE FROM THE SOS+SONC CONE - A SOS+SONC ANALOGUE TO HILBERT'S 1888 THEOREM

As in *Hilbert 1888* for the SOS case and the theorem of Section 2.5 for the SONC case, it is of interest to find separation results, which classify whenever a given inner approximation of the PSD cone is proper or not. Therefore, we show in Theorem 1 an analogue to *Hilbert 1888* for the SOS+SONC case. The following statement was already proven in a non constructive way in (Averkov, 2019, Corollary 2.17). As a contribution we present an alternative proof by constructing appropriate polynomials in the two basic cases $((n, 2d) \in \{(3, 6), (4, 4)\})$ and scaling them to higher dimensional and number of variables cases.

Theorem 1. It holds $(\Sigma + C)_{n,2d} = P_{n,2d}$ if and only if $n = 2$ or $2d = 2$ or $(n, 2d) = (3, 4)$.

As a first step, we show that it suffices to consider the two elementary cases of ternary sextics and quarternary quartics, i.e. $(n, 2d) \in \{(3, 6), (4, 4)\}$.

Lemma 2. If $(\Sigma+C)_{3,6} \subsetneq P_{3,6}$ and $(\Sigma+C)_{4,4} \subsetneq P_{4,4}$ then $(\Sigma_{n,k}+C)_{n,k} \subsetneq P_{n,k}$ for all $n \geq 3, k \geq 4$ and $(n,k) \neq (3,4)$ (k even).

Proof. We show equivalently: $P_{3,6}\backslash(\Sigma + C)_{3,6} \neq \emptyset$ and $P_{4,4}\backslash(\Sigma + C)_{4,4} \neq \emptyset$ imply $P_{n,k}\backslash(\Sigma + C)_{n,k} \neq \emptyset$ for all $n \geq 3, k \geq 4$ and $(n,k) \neq (3,4)$ where k is even.

Claim 1: $f \in P_{n,k}\backslash(\Sigma + C)_{n,k}$ implies for all $m \in \mathbb{N}$, $f \in P_{n+m,k}\backslash(\Sigma + C)_{n+m,k}$.

The case $m = 1$ can be seen easily, since SOS and SONC polynomials in $H_{n+1,k}$ stay SOS and SONC, respectively, after plugging in $\mathrm{x}_{n+1} = 0$. The general case follows inductively.

Claim 2: $f \in P_{n,k}\backslash(\Sigma + C)_{n,k}$ implies for all $\ell \in \mathbb{N}$ $\mathrm{x}_1^{2\ell} f \in P_{n,k+2\ell}\backslash(\Sigma + C)_{n,k+2\ell}$.

Consider $\ell = 1$ and let $f \in P_{n,k}\backslash(\Sigma + C)_{n,k}$ be arbitrary. Assume that $\mathrm{x}_1^2 f \in (\Sigma + C)_{n,k+2}$, i.e. there is a decomposition $\mathrm{x}_1^2 f = f_{\mathrm{SOS}} + f_{\mathrm{SONC}}$ with f_{SOS} and f_{SONC} being SOS and SONC polynomials in $H_{n,k+2}$, respectively. Since the left hand side vanishes at $\mathrm{x}_1 = 0$, the right hand side must vanish at $\mathrm{x}_1 = 0$ as well. Since $f_{\mathrm{SOS}}, f_{\mathrm{SONC}}$ are PSD, we obtain $f_{\mathrm{SOS}}(0, \mathrm{x}_2, \ldots, \mathrm{x}_n) = 0 = f_{\mathrm{SONC}}(0, \mathrm{x}_2, \ldots, \mathrm{x}_n)$. Hence $\mathrm{x}_1 \mid f_{\mathrm{SOS}}, f_{\mathrm{SONC}}$, i.e. $f_{\mathrm{SOS}} = \mathrm{x}_1 \cdot f_1, f_{\mathrm{SONC}} = \mathrm{x}_1 \cdot f_2$ for some $f_1, f_2 \in H_{n,k+1}$.

Write $f_{\mathrm{SOS}} = \sum_{i=1}^s g_i^2$, $f_{\mathrm{SONC}} = \sum_{j=1}^t h_j, s, t \in \mathbb{N}$ s.t. $g_i \in H_{n,k/2+1}$ and the h_j are nonnegative circuit polynomials. Similary as above, PSDness of g_i^2 and h_j yields $\mathrm{x}_1 \mid g_i^2, h_j$. For the SOS part, we immediately obtain $\mathrm{x}_1 \mid g_i$, i.e. $\mathrm{x}_1^2 \mid f_{\mathrm{SOS}}$. Remains to show $\mathrm{x}_1^2 \mid h_j$ for all j. However, this follows since all monomials of h_j must be divisible by x_1, all vertex monomials of x_1 are even lattice points and the only interior monomial is a convex combination of the vertex monomials. Hence we have $\mathrm{x}_1^2 \mid f_{\mathrm{SOS}}, f_{\mathrm{SONC}}$, which yields that $f = f_{\mathrm{SOS}}/\mathrm{x}_1^2 + f_{\mathrm{SONC}}/\mathrm{x}_1^2$ would be a SOS+SONC decomposition in $H_{n,k}$, a contradiction. This shows Claim 2 for $\ell = 1$. The general case follows inductively.

Combining Claim 1 and Claim 2 shows the Lemma.

Next, we cover the two elementary cases where $(n, 2d) \in \{(3,6),(4,4)\}$. Therefore, we show that the two Robinson forms from Robinson (1969) are indeed PSD but not SOS+SONC.

Lemma 3. It holds

$$\begin{aligned} R_1(\mathrm{x},\mathrm{y},\mathrm{z}) =& \mathrm{x}^6 + \mathrm{y}^6 + \mathrm{z}^6 - \left(\mathrm{x}^4\mathrm{y}^2 + \mathrm{x}^4\mathrm{z}^2 + \mathrm{y}^4\mathrm{x}^2 + \mathrm{y}^4\mathrm{z}^2\right. \\ &\left. + \mathrm{z}^4\mathrm{x}^2 + \mathrm{z}^4\mathrm{y}^2\right) + 3\mathrm{x}^2\mathrm{y}^2\mathrm{z}^2 \in P_{3,6}\backslash(\Sigma + C)_{3,6} \end{aligned}$$

and hence $P_{3,6} \neq (\Sigma + C)_{3,6}$.

Proof. Step 1: Assume that $R_1 \in (\Sigma + C)_{3,6}$ was SOS+SONC. Choose an SOS polynomial $f_{\mathrm{SOS}} \in \Sigma_{3,6}$ such that $R_1 - f_{\mathrm{SOS}} \in C_{3,6}$ is a SONC polynomial. Without loss of generality we can assume that $R_1 - f_{\mathrm{SOS}}$ decomposes into nonnegative circuit polynomials, which are not SOS. Since $R_1 = f_{\mathrm{SOS}} + (R_1 - f_{\mathrm{SOS}})$ is a decomposition into PSD polynomials, $\mathrm{New}(f_{\mathrm{SOS}})$, $\mathrm{New}(R_1 - f_{\mathrm{SOS}}) \subseteq \mathrm{New}(R_1)$ must hold (cf. (Reznick, 1978, Theorem 1)).

Step 2: By *Hilbert 1888*, we know that every PSD bivariate form is SOS, i.e. $P_{2,2d} = \Sigma_{2,2d}$ for all $d \in \mathbb{N}$. Now consider e.g. the monomial $m = \mathrm{x}^4\mathrm{y}^2$. Assume that the SONC part $R_1 - f_{\mathrm{SOS}}$ contains a nonnegative circuit polynomial h having m as interior monomial, i.e. $h_{(4,2,0)'} \lneq 0$. But then, since m is in the interior of $\mathrm{New}(h)$, the circuit polynomial h must clearly be bivariate, i.e. $h \in C_{2,6} \subseteq P_{2,6} = \Sigma_{2,6}$. This contradicts our assumption that no nonnegative circuit polynomial in $R_1 - f_{\mathrm{SOS}}$ is also SOS. Hence, m cannot be an interior monomial of any nonnegative circuit polynomials in the decomposition of $R_1 - f_{\mathrm{SOS}}$, which shows that $(R_1 - f_{\mathrm{SOS}})_{(4,2,0)'} \geq 0$ and equivalently $(f_{\mathrm{SOS}})_{(4,2,0)'} \leq -1$. Analogous argumentation shows

$$\left.\begin{matrix}(f_{\mathrm{SOS}})_{(4,2,0)'},\ (f_{\mathrm{SOS}})_{(4,0,2)'},\ (f_{\mathrm{SOS}})_{(2,4,0)'},\\ (f_{\mathrm{SOS}})_{(2,0,4)'},\ (f_{\mathrm{SOS}})_{(0,4,2)'},\ (f_{\mathrm{SOS}})_{(0,2,4)'}\end{matrix}\right\} \leq -1 \quad (1)$$

This yields in particular $\mathrm{x}^6, \mathrm{y}^6, \mathrm{z}^6 \in \mathrm{New}(f_{\mathrm{SOS}})$ and hence $\mathrm{New}(f_{\mathrm{SOS}}) = \mathrm{New}(R_1)$.

Step 3: Write $f_{\mathrm{SOS}} = \sum_{i=1}^s g_i^2$ s.t. $g_i \in H_{3,3}$, $s \in \mathbb{N}$. Since $\mathrm{New}(g_i) \subseteq \frac{1}{2}\mathrm{New}(f_{\mathrm{SOS}}) = \frac{1}{2}\mathrm{New}(R_1) = \mathrm{conv}(\mathrm{x}^3, \mathrm{y}^3, \mathrm{z}^3)$, all possible exponents of the g_i's are given by

$$\begin{aligned} &\alpha(1) = (3,0,0)',\ \alpha(2) = (2,1,0)',\ \alpha(3) = (2,0,1')),\\ &\alpha(4) = (1,2,0)',\ \alpha(5) = (1,1,1)',\ \alpha(6) = (1,0,2)',\\ &\alpha(7) = (0,3,0)',\ \alpha(8) = (0,2,1)',\ \alpha(9) = (0,1,2)^2,\\ &\alpha(10) = (0,0,3)' \end{aligned}$$

and we can write $g_i = \sum_{j=1}^{10} g_{ij}\mathrm{x}^{\alpha(j)}$ $(i = 1, \ldots, s)$, for some $g_{ij} \in \mathbb{R}$.

Step 4: By (1) and the decomposition of $f_{\mathrm{SOS}} = \sum_{i=1}^s g_i^2$, we know $-1 \geq (f_{\mathrm{SOS}})_{(4,2,0)'} = \sum_{i=1}^s g_{i,2}^2 + \sum_{i=1}^s 2g_{i,1}g_{i,4}$. Hence, using Young's inequality we obtain

$$\begin{aligned} -1 \geq (f_{\mathrm{SOS}})_{(4,2,0)'} &= \sum_{i=1}^s g_{i,2}^2 + \sum_{i=1}^s 2g_{i,1}g_{i,4} \\ &\geq \sum_{i=1}^s g_{i,2}^2 - \sum_{i=1}^s 2\,|g_{i,1}| \cdot |g_{i,4}| \\ &\geq \sum_{i=1}^s g_{i,2}^2 - \sum_{i=1}^s g_{i,1}^2 - \sum_{i=1}^s g_{i,4}^2 \\ &\geq -1 + \sum_{i=1}^s g_{i,2}^2 - \sum_{i=1}^s g_{i,4}^2, \end{aligned} \quad (2)$$

where we used that $(f_{\mathrm{SOS}})_{(6,0,0)'} = \sum_{i=1}^s g_{i,1}^2 \leq 1$ must hold. Rearranging (2) yields $\sum_{i=1}^s g_{i,2}^2 \leq \sum_{i=1}^s g_{i,4}^2$.

Analogous argumentation shows that e.g.

$$\begin{aligned} -1 \geq (f_{\mathrm{SOS}})_{(2,4,0)} &= \sum_{i=1}^s g_{i,4}^2 + \sum_{i=1}^s 2g_{i,7}g_{i,2} \\ &\geq \cdots \geq -1 + \sum_{i=1}^s g_{i,4}^2 - \sum_{i=1}^s g_{i,2}^2 \end{aligned}$$

and therefore $\sum_{i=1}^s g_{i,2}^2 \geq \sum_{i=1}^s g_{i,4}^2$. We finally obtain the equality $\sum_{i=1}^s g_{i,2}^2 = \sum_{i=1}^s g_{i,4}^2$. Hence, equality must hold everywhere in (2), which shows:

(III) $(f_{\mathrm{SOS}})'_{(6,0,0)} = \sum_{i=1}^s g_{i,1}^2 = -1$.
(IV) By Young's Inequality: $g_{i,1} \neq 0$ if and only if $g_{i,4} \neq 0$ and in this case $|g_{i,1}| = |g_{i,4}|$ $(i \in [s])$.
(V) $\mathrm{sign}(g_{i,1}) = -\mathrm{sign}(g_{i,4})$ $(i \in [s])$.
(VI) $(f_{\mathrm{SOS}})_{(4,2,0)'} = -1$.

Clearly, similar observations as in (III)-(V) can be made for all pairs $(g_{i,r}, g_{i,s})$ s.t. $(r,s) \in \{(1,4),(1,6),(7,2),(7,9),(10,3),(10,8)\}$. In addition, as in (VI), we obtain for the other coefficients as in (1):

$$(f_{\text{SOS}})_{(4,2,0)'} = (f_{\text{SOS}})_{(4,0,2)'} = \ldots = -1. \tag{3}$$

Step 5: By (3), we have

$$(R_1 - f_{\text{SOS}})_{(4,2,0)'} = (R_1 - f_{\text{SOS}})_{(4,0,2)'} = \ldots = 0.$$

Furthermore, (III) for all possible coefficients leads to

$$\begin{aligned}&(R_1 - f_{\text{SOS}})_{(6,0,0)'} = (R_1 - f_{\text{SOS}})_{(0,6,0)'}\\ =&(R_1 - f_{\text{SOS}})_{(0,0,6)'} = 0.\end{aligned}$$

To sum up, we now have x^6, y^6, z^6, $\mathrm{x}^4\mathrm{y}^2$, $\mathrm{x}^4\mathrm{z}^2, \mathrm{x}^2\mathrm{y}^4$, $\mathrm{x}^2\mathrm{z}^4$, $\mathrm{y}^4\mathrm{z}^2, \mathrm{y}^2\mathrm{z}^4 \notin \operatorname{supp}(R_1 - f_{\text{SOS}})$. However, the form $R_1 - f_{\text{SOS}}$ is SONC and in particular PSD. Hence all vertices in $V(R_1 - f_{\text{SOS}})$ are even. Since $\operatorname{New}(R_1 - f_{\text{SOS}}) \subseteq \operatorname{New}(R_1)$, the only possible lattice point left is $\mathrm{x}^2\mathrm{y}^2\mathrm{z}^2$. Hence, $\operatorname{New}(R_1 - f_{\text{SOS}}) \subseteq \operatorname{conv}(\mathrm{x}^2\mathrm{y}^2\mathrm{z}^2) = \{\mathrm{x}^2\mathrm{y}^2\mathrm{z}^2\}$ must hold and $R_1 - f_{\text{SOS}}$ would be SOS, which is a contradiction.

For the quarternary quartics case, we can argue similarly.

Lemma 4. It holds $P_{4,4} \neq (\Sigma + C)_{4,4}$. More precisely, we have

$$\begin{aligned}R_2(\mathrm{x},\mathrm{y},\mathrm{z},\mathrm{w}) = &\,\mathrm{x}^2(\mathrm{x}-\mathrm{w})^2 + \mathrm{y}^2(\mathrm{y}-\mathrm{w})^2 + \mathrm{z}^2(\mathrm{z}-\mathrm{w})^2\\ &+ 2\mathrm{xyz}(\mathrm{x}+\mathrm{y}+\mathrm{z}-2\mathrm{w}) \in P_{4,4}\backslash(\Sigma + C)_{4,4}.\end{aligned}$$

Proof. The proof follows an analogous argumentation as in Lemma 3. By *Hilbert 1888*, we can without loss of generality choose $f_{\text{SOS}} \in \Sigma_{4,4}$ s.t. $R_2 - f_{\text{SOS}} \in C_{4,4}$ is SONC and does not contain any nonnegative circuit polynomial in three variables in its decomposition. It can be deduced that $\operatorname{New}(f_{\text{SOS}}) = \operatorname{New}(R_2)$ must hold.

Further, using Young's inequality for the coefficients of $\mathrm{x}^3\mathrm{w}$, $\mathrm{y}^3\mathrm{w}$, $\mathrm{z}^3\mathrm{w}$ it can be seen that

$$\mathrm{x}^4,\ \mathrm{x}^2\mathrm{w}^2,\ \mathrm{y}^4,\ \mathrm{y}^2\mathrm{w}^2,\ \mathrm{z}^4,\ \mathrm{z}^2\mathrm{w}^2 \notin \operatorname{supp}(R_2 - f_{\text{SOS}}).$$

Hence, the only even exponents in $\operatorname{supp}(R_2)$ left as possible lattice points for $R_2 - f_{\text{SOS}}$ are $\mathrm{x}^2\mathrm{y}^2$, $\mathrm{x}^2\mathrm{z}^2$, $\mathrm{y}^2\mathrm{z}^2$. However, this means that $R_2 - f_{\text{SOS}}$ is a PSD form in three variables of degree four, which is SOS by *Hilbert 1888*. Hence, R_2 is SOS as well, which is a contradiction. For this reason, R_2 cannot be SOS+SONC.

We are now able to prove Theorem 1.

Proof. [Theorem 1.] "$\Leftarrow$" is clear by *Hilbert 1888*.

"$\Rightarrow$": The Robinson polynomials from Lemma 3 and Lemma 4 are examples of polynomials in $P_{3,6}\backslash(\Sigma + C)_{3,6}$ and $P_{4,4}\backslash(\Sigma + C)_{4,4}$, respectively. Hence, the claim follows directly from Lemma 2.

4. SEPARATING THE SOS+SONC CONE FROM THE SOS AND SONC CONE

In this section, we present a theorem which shows that the SOS+SONC cone is a proper cone extension of both the SOS and the SONC cones for all $(n,2d) \geq (3,4), (n,2d) \neq (3,4)$. This shows that for all nontrivial $(n,2d)$, the SOS+SONC cone is a better inner approximation of the PSD cone than the single SOS and SONC cones.

Theorem 5. For all $(n,2d) \geq (3,4), (n,2d) \neq (3,4)$ it holds

$$(\Sigma + C)_{n,2d} \not\subseteq (\Sigma_{n,2d} \cup C_{n,2d})\,.$$

Proof. Similarly as in Lemma 2 one can show that it suffices to consider the cases $(n,2d) \in \{(3,6),(4,4)\}$. Hence, the claim follows by constructing explicit examples for the two elementary cases. Indeed, we have e.g.

$$\begin{aligned}f_1 =&\,\mathrm{x}^4\mathrm{y}^2 + \mathrm{x}^2\mathrm{y}^4 + \mathrm{z}^6 - 3\mathrm{x}^2\mathrm{y}^2\mathrm{z}^2\\ &+ {}^1\!/_2 \cdot (\mathrm{z}^3 + 2\mathrm{xyz} + \mathrm{x}^2\mathrm{y})^2 \in (\Sigma + C)_{3,6}\backslash(\Sigma \cup C)_{3,6},\\ f_2 =&\,\mathrm{x}^2\mathrm{y}^2 + \mathrm{x}^2\mathrm{z}^2 + \mathrm{y}^2\mathrm{z}^2 + \mathrm{w}^4 - 4\mathrm{wxyz}\\ &+ (\mathrm{xy} + \mathrm{xz} + \mathrm{yz})^2 + \mathrm{w}^4 \in (\Sigma + C)_{4,4}\backslash(\Sigma \cup C)_{4,4}.\end{aligned}$$

5. CONCLUSION

Combining Theorem 1 and Theorem 5 we have shown that for all non Hilbert cases $(n,2d) \geq (3,4)$, $(n,2d) \neq (3,4)$, it holds $(\Sigma_{n,2d} \cup C_{n,2d}) \subsetneq (\Sigma + C)_{n,2d} \subsetneq P_{n,2d}$.

We presented explicit examples R_1, R_2 and f_1, f_2 showing the inequalities for the two basic cases and demonstrated how two scale them to arbitrary cases. In terms of polynomial optimization, this shows that for all mentioned cases of n and $2d$, there are PSD polynomial which can be handled by the SOS+SONC cone but neither the SOS nor the SONC cone themselves. On the other hand, there are polynomials which are not classifiable as being PSD by SOS+SONC. It remains to find an efficient way to actually decide membership to the combined SOS+SONC cone.

REFERENCES

Averkov, G. (2019). Optimal size of linear matrix inequalities in semidefinite approaches to polynomial optimization. *SIAM J. Appl. Algebra Geometry*, 3, 128–151. doi: 10.1137/18M1201342.

Dressler, M. (2018). *Sums of nonnegative circuit polynomials: geometry and optimization*. Ph.D. thesis, Goethe-Universität Frankfurt am Main.

Dressler, M., Heuer, J., Naumann, H., and de Wolff, T. (2020). Global optimization via the dual sonc cone and linear programming. In *ISSAC '20: Proceedings of the 45th International Symposium on Symbolic and Algebraic Computation*, 138–145. doi: 10.1145/3373207.3404043.

Hilbert, D. (1888). *Über die Darstellung definiter Formen als Summe von Formenquadraten*, volume 32. doi: 10.1007/BF01443605.

Iliman, S. and de Wolff, T. (2016). Amoebas, nonnegative polynomials and sums of squares supported on circuits. *Res. Math. Sci.*, 3. doi:10.1186/s40687-016-0052-2. 3:9.

Lasserre, J. (2009). *Moments, Positive Polynomials and Their Applications*, volume 1 of *Imperial College Press Optimization Series*. Imperial College Press. doi: 10.1142/p665.

Reznick, B. (1978). Extremal PSD forms with few terms. *Duke Math. J.*, 45(2). doi:10.1215/S0012-7094-78-04519-2.

Robinson, R. (1969). Some definite polynomials which are not sums of squares of real polynomials. *Notices Amer. Math. Soc.*, 16(3), 554.

Wang, J. and Magron, V. (2020). A second order cone characterization for sums of nonnegative circuits. In *ISSAC '20: Proceedings of the 45th International Symposium on Symbolic and Algebraic Computation*, 450–457. doi:10.1145/3373207.3404033.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Compatible snapshot-based model reduction of a nonlinear port-Hamiltonian PDE on networks ⋆

Björn Liljegren-Sailer * **Nicole Marheineke** *

* *Trier University, 54286 Trier, Germany (e-mail: {bjoern.sailer, marheineke}@uni-trier.de).*

Abstract: This contribution is on the construction of structure-preserving, online-efficient reduced models for a class of nonlinear partial differential equations on networks, which inherit a port-Hamiltonian structure. The flow problem finds broad application, e.g., in the context of gas distribution networks. We propose a snapshot-based reduction approach that consists of a mixed variational Galerkin approximation combined with quadrature-type complexity reduction. Its main feature is that certain compatibility conditions are assured during the training phase, which make our approach structure-preserving. The resulting reduced models are locally mass conservative and inherit an energy-bound and port-Hamiltonian structure. We demonstrate the applicability and good stability properties of our approach using the example of the Euler equations on networks.

Keywords: port-Hamiltonian systems; structure-preserving scheme; systems on graphs; model reduction; Legendre transformation; network systems; Euler equations.

1. INTRODUCTION

Structure-preserving approximation is an active research area. By preserving or mimicking relevant geometric structures such as, e.g., conservation laws, dissipative relations, or symplecticities, unphysical solution behavior and numerical instabilities can be avoided in many cases. The model problem in this contribution describes nonlinear flows on networks and has a port-Hamiltonian structure. It finds, e.g., application in the context of gas network systems or electric transmission lines. The network is assumed to be described by a directed graph. Each edge ω of the graph can be identified with an interval. Given a strictly convex smooth function $h : \mathbb{R}^2 \to \mathbb{R}$ and a non-negative function $\tilde{r} : \mathbb{R}^2 \to \mathbb{R}$, the edgewise states $\underline{\mathbf{z}}^\omega = [z_1^\omega; z_2^\omega] : [0,T] \times \omega \to \mathbb{R}^2$ are governed by

$$\begin{aligned} \partial_t z_1^\omega(t,x) &= -\partial_x \nabla_2 h(\underline{\mathbf{z}}^\omega(t,x)), \\ \partial_t z_2^\omega(t,x) &= -\partial_x \nabla_1 h(\underline{\mathbf{z}}^\omega(t,x)) \\ &\quad - \tilde{r}(\underline{\mathbf{z}}^\omega(t,x)) \nabla_2 h(\underline{\mathbf{z}}^\omega(t,x)), \end{aligned} \tag{1}$$

whereby $\nabla_i h(\underline{\mathbf{z}}^\omega(t,x)) = \partial_{z_i} h([z_1; z_2])_{|[z_1;z_2]=\underline{\mathbf{z}}^\omega(t,x)}$ for $i = 1, 2$. The expressions $\mathcal{M}(\underline{\mathbf{z}}) = \sum_{\omega\in\mathcal{E}} \int_\omega z_1^\omega dx$ and $\tilde{\mathcal{H}}(\underline{\mathbf{z}}) = \sum_{\omega\in\mathcal{E}} \int_\omega h(\underline{\mathbf{z}}^\omega) dx$, with $\mathcal{E}$ the set of all edges, represent the total mass and the Hamiltonian of the system. Fundamental properties of the hyperbolic model problem are that, under appropriate coupling conditions on the edgewise equations, conservation of mass and dissipation of the Hamiltonian (energy dissipation) hold up to the exchange with the boundary. Moreover, the convective terms can be related to a certain skew symmetric geometric structure.

Following the approach by Liljegren-Sailer and Marheineke (2021), we aim for a structure-preserving approximation procedure for the flow problem on network, which can be used for the discretization by finite elements, as well as the subsequent snapshot-based model reduction. The model reduction consists in general of a projection-based step and an additional complexity reduction of the nonlinearities employing a quadrature-based approach in our case. In this brief, we summarize and rephrase the results from Liljegren-Sailer and Marheineke (2021) in the abstract port-Hamiltonian setting.

For ease of presentation, we consider the case of a single edge in the remaining sections, i.e., without loss of generality the spatial domain $\Omega = [0, \ell]$ with $\ell > 0$. For the generalization to the network case and the choice of coupling-conditions, we refer to Liljegren-Sailer and Marheineke (2020).

2. AN EXEMPLARY HIERARCHY OF MODELS

Flow models of different complexity are covered by the problem (1), e.g., the following hierarchy used for the modeling of gas distribution networks, cf., Mindt et al. (2019).

- *Isothermal Euler equations:* The equations for density ρ and velocity v for $x \in \Omega$ and $t \geq 0$ read

$$\begin{aligned} \partial_t \rho &= -\partial_x(\rho v) \\ \partial_t(\rho v) &= -\partial_x(\rho v^2 + \hat{p}(\rho)) - c_f \frac{|\rho v|}{\rho} \rho v \end{aligned}$$

⋆ The support of the German Federal Ministry of Education and Research (BMBF) via the project *EiFer* is acknowledged. Moreover, we thank for the support of the DFG research training group 2126 on algorithmic optimization.

with pressure $p(\rho) = RT\frac{\rho}{1-RT\alpha\rho}$ (constants $RT > 0$, $\alpha < 0$). To rewrite the system as in (1), we define $\underline{\mathbf{z}} = [\rho; v]$, $\tilde{r}(\underline{\mathbf{z}}) = c_f\frac{|v|}{\rho}$ and

$$h(\underline{\mathbf{z}}) = \rho\frac{v^2}{2} + RT\rho\log\left(\rho_{sc}\frac{1-RT\alpha\rho}{\rho}\right).$$

- *Simplified isothermal Euler equations:* These equations result from the isothermal Euler equations by neglecting the term $\partial_x(\rho v^2)$. Reformulation (1) is obtained for $m = \rho v$, $\tilde{r}(\underline{\mathbf{z}}) = c_f\frac{|m|}{\rho}$ and

$$h(\underline{\mathbf{z}}) = \frac{1}{2}m^2 + \frac{-1}{RT\alpha^2}\left(\log\left(1-RT\alpha\rho\right) + RT\alpha\rho\right).$$

- *Damped wave equation:* A damped wave equation is obtained for

$$h(\underline{\mathbf{z}}) = 1/2(z_1^2+z_2^2), \quad \tilde{r}(\underline{\mathbf{z}}) > 0 \quad \text{for all } \underline{\mathbf{z}}.$$

The system is linear, when $\tilde{r}$ is chosen to be constant.

Note that in contrast to the isothermal Euler equations, the simplified version inherits a Hamiltonian density which is separable into $h(\underline{\mathbf{z}}) = h_1(z_1) + h_2(z_2)$ with quadratic h_2. For the damped wave equation, the Hamiltonian density is a fully quadratic function and $\underline{\mathbf{z}} = \nabla_{\underline{\mathbf{z}}}h(\underline{\mathbf{z}})$.

3. VARIATIONAL PRINCIPLE

Let the domain $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2 \subset \mathbb{R}^2$ of the Hamiltonian density h be open and convex. Then the so-called partial Legendre transformation g for $\quad = [a_1; a_2]$ is given by

$$g(\ ) = \sup_{z_2 \in \{\bar{z}_2 : [a_1;\bar{z}_2]\in\mathbb{S}\}} a_2 \cdot z_2 - h([a_1; z_2]).$$

It inherits the smoothness and strict convexity w.r.t. the second argument from h, cf., Rockafellar and Wets (1998). A certain variable transformation is induced by the Legendre transformation: Let for $\mathbf{z} \in \mathbb{S}$, $\quad = [z_1; \nabla_2 h(\mathbf{z})]$. Then it follows $\nabla_1 h(\mathbf{z}) = -\nabla_1 g(\ )$ and $z_2 = \nabla_2 g(\ )$, i.e., $a_2 \mapsto \nabla_2 g([z_1; a_2])$ is the inverse function of $z_2 \mapsto \nabla_2 h([z_1; z_2])$.

Thus, when $\underline{\mathbf{z}}(t,x) = [z_1(t,x); z_2(t,x)]$ is a smooth function, for which (1) holds for $t \geq 0$ and $x \in \Omega$, the function $_(t,x) = [a_1(t,x); a_2(t,x)] = [z_1(t,x); \nabla_2 h(\underline{\mathbf{z}}(t,x))]$ fulfills

$$\begin{aligned} \partial_t a_1(t,x) &= -\partial_x \nabla_2 a_2(t,x)), \\ \partial_t \nabla_2 g(a_2(t,x)) &= \partial_x \nabla_1 g(_(t,x)) - r(_(t,x))a_2(t,x) \end{aligned} \tag{2}$$

for a non-negative mapping r. The system can be closed by, e.g., the initial conditions $_(0,x) = _0(x)$ and boundary-conditions $-[\nabla_1 g(_(t,0); \nabla_1 g(_(t,\ell))] = \quad (t)$, $t \geq 0$ for given $_0 : \Omega \to \mathbb{R}^2$ and $\quad : [0,T] \to \mathbb{R}^2$.

Let the function spaces $\mathcal{L}^2(\Omega)$ and $\mathcal{H}^1(\Omega)$ denote the Sobolev space of square integrable functions on Ω and the Sobolev space with additionally square integrable weak derivatives, respectively. The $\mathcal{L}^2$-scalar product is written as $\langle\cdot,\cdot\rangle$ and the boundary terms for $b \in \mathbb{H}^1(\Omega)$ are denoted as $b[0]$ and $b[\ell]$. Accordingly, we define the boundary operator $\mathbf{T} : \mathbb{H}^1(\Omega) \to \mathbb{R}^2$ as $\mathbf{T}b = -[b[0]; b[\ell]]$. For any $_ \in \mathcal{C}^1([0,T], \mathcal{C}^1(\Omega) \times \mathcal{C}^1(\Omega))$ fulfilling (2) the variational principle

$$\begin{aligned} \langle \partial_t a_1(t), b_1\rangle &= -\langle \partial_x \nabla_2 a_2(t)), b_1\rangle \\ \langle \partial_t \nabla_2 g(a_2(t)), b_2\rangle &= -\langle \nabla_1 g(_(t)), \partial_x b_2\rangle \\ &\quad - \langle r(_(t))a_2(t), b_2\rangle + \quad (t)\cdot\mathbf{T}b_2. \end{aligned}$$

for $b_1 \in \mathcal{L}^2(\Omega)$, $b_2 \in \mathcal{H}^1(\Omega)$ holds, as one shows by multiplying (2) with the test functions b_1, b_2 and using integration by parts in the second equation. Note that here the solution is interpreted as a function in time with values in a function space.

4. APPROXIMATION ANSATZ

As we showed in Liljegren-Sailer and Marheineke (2020), a class of structure-preserving approximations is obtained by Galerkin approximation with compatible spaces that can be supplemented by a quadrature-type complexity reduction. The latter allows for sparse approximations of the nonlinearities and is required for reduced models to be online-efficient.

Assumption 1. (Compatibility of spaces). Let $\mathcal{V} = \mathcal{Q} \times \mathcal{W} \subset \mathcal{L}^2(\Omega) \times \mathcal{H}^1(\Omega)$ be a finite dimensional subspace fulfilling the compatibility conditions

A1) $\mathcal{Q} = \{\xi : \text{It exists } \zeta \in \mathcal{W} \text{ with } \partial_x\zeta = \xi\}$,
A2) $\{b_2 \in \mathcal{H}^1(\Omega) : \partial_x b_2 = 0\} \subset \mathcal{W}$.

Assumption 2. (Compatibility of scalar product). Let the bilinear form $\langle\cdot,\cdot\rangle_* : \mathcal{L}^2(\Omega) \times \mathcal{L}^2(\Omega) \to \mathbb{R}$ be such that the following holds:

A1) For a constant $\tilde{C} \geq 1$ and $||b||_* = \sqrt{\langle b, b\rangle_*}$, it holds $\tilde{C}^{-1}||b||_* \leq ||b|| \leq \tilde{C}||b||_*$ for all $b \in \mathcal{Q} \cup \mathcal{W}$.
A2) For any $f \in \mathcal{C}(\Omega)$ with $f \geq 0$ it holds $\langle f, 1\rangle_* \geq 0$.

The posed assumption assures that $\langle\cdot,\cdot\rangle_*$ is a scalar product on $\mathcal{Q}$ and $\mathcal{W}$. Note that the $\mathcal{L}^2$-product is particularly one possible choice, which results in a model without complexity reduction.

For approximation spaces and bilinear products as in the assumptions, our abstract approximation ansatz reads as follows: Given $_0 \in \mathcal{Q} \times \mathcal{W}$ and $\quad : [0,T] \to \mathbb{R}^2$, find $_ = [a_1; a_2] \in \mathcal{C}^1([0,T]; \mathcal{Q} \times \mathcal{W})$ with $_(0) = _0$, and

$$\begin{aligned} \langle \partial_t a_1(t), b_1\rangle &= -\langle \partial_x \nabla_2 a_2(t)), b_1\rangle \\ \langle \partial_t \nabla_2 g(a_2(t)), b_2\rangle_* &= -\langle \nabla_1 g(_(t)), \partial_x b_2\rangle_* \\ &\quad - \langle r(_(t))a_2(t), b_2\rangle_* + \quad (t)\cdot\mathbf{T}b_2. \end{aligned}$$

for $b_1 \in \mathcal{Q}$, $b_2 \in \mathcal{W}$.

The proposed approximations can be shown to be of port-Hamiltonian form with the Hamiltonian defined as $\mathcal{H}_*(_(t)) = \langle \nabla_2 g(_(t))a_2(t) - g(_(t)), 1\rangle_*$. Moreover, they inherit energy dissipation and mass conservation up to the exchange with the boundary,

$$\begin{aligned} \frac{d}{dt}\mathcal{H}_*(_(t)) &= -\langle r(_(t)), a_2(t)^2\rangle_* + \quad (t)\cdot\mathbf{T}a_2(t) \\ &\leq \quad (t)\cdot\mathbf{T}a_2(t) \\ \frac{d}{dt}\int_\Omega a_1(t)dx &= a_2(t)[0] + a_2(t)[\ell]. \end{aligned}$$

5. APPLICATION

The presented approximation framework can be realized by mixed finite elements but also can be used for the construction of online-efficient reduced models. The numerical results in this section focus on the reduction steps, i.e., the model order- and complexity-reduction. We showcase the improved performance when regarding the compatibility conditions as opposed to using standard (non-compatible) methods. We refer to Liljegren-Sailer and Marheineke

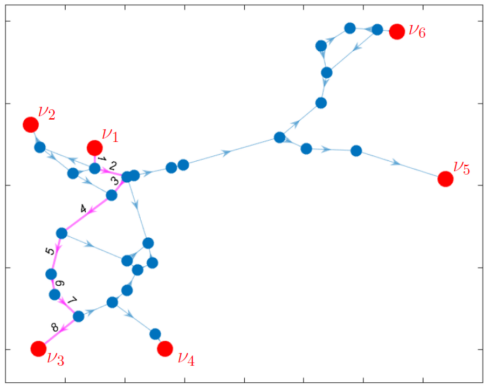


Fig. 1. Topology of network with boundary nodes ν_i, $i = 1, ..., 6$ (red circles).

(2021) for the algorithmic treatment of the compatibility conditions, the complete set of parameters and more numerical results. The test case employs the isothermal Euler equations in a regime relevant for gas distribution networks and the topology visualized in Fig. 1. As starting point for all reduction methods, a mixed finite element discretization with a system dimension of $N = 10\,156$ is used.

We simulate the high dimensional system using the boundary conditions for $t \in [0, 5t_\star]$ (reference values marked by stars, e.g., time $t_\star = 1[h]$),

$$\rho(t,\nu_1) = (65 + u(t/t_\star))\,\rho_\star, \qquad \rho(t,\nu_2) = (50 + u(t/t_\star))\,\rho_\star,$$
$$\rho(t,\nu_4) = (60 - u(t/t_\star))\,\rho_\star, \qquad \rho(t,\nu_5) = 60\,\rho_\star,$$
$$\rho(t,\nu_6) = 45\,\rho_\star, \qquad Am(t,\nu_3) = -100\,(Am)_\star,$$

at the six boundary nodes ν_i, $i = 1, ..., 6$ with the input profile $u = u_A$ (training case) and $u = u_B$ (testing case),

$$u_A(t) = 6\exp\left(-\frac{3}{2}t\right) + 4\cos\left(\frac{\pi}{2}t\right) + \frac{3}{2}\sin(10\pi\,t)$$

$$u_B(t) = 8t^3\exp(-t) - 4\,(t-2)\,f(3t).$$

with $f(t) = 1 - |(t \bmod 2) - 1|$. As initial condition the respective stationary solution for $t = 0$ is chosen. Reduced models without and with complexity reduction are obtained using snapshots from the training case. The fidelity of the resulting models are compared in Fig. 2 and Fig. 3 for varying dimensions of in the model order- and complexity-reduction. As is observed, our proposed schemes clearly outperform the non-structure-preserving conventional reduced models in the test case.

REFERENCES

Liljegren-Sailer, B. and Marheineke, N. (2020). On port-Hamiltonian approximation of a nonlinear flow problem on networks. arXiv e-prints 2009.11216.

Liljegren-Sailer, B. and Marheineke, N. (2021). On snapshot-based model reduction under compatibility conditions for a nonlinear flow problem on networks. arXiv e-prints 2110.04777.

Mindt, P., Lang, J., and Domschke, P. (2019). Entropy-preserving coupling of hierarchical gas models. *SIAM J. Math. Anal.*, 51(6), 4754–4775. doi: 10.1137/19M1240034.

Rockafellar, R. and Wets, R. (1998). *Variational Analysis*. Springer.

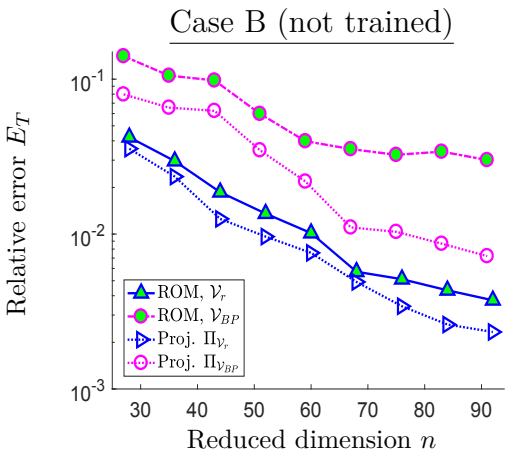


Fig. 2. Errors of ROMs and orthogonal projection onto the respective reduction space (Proj. $\Pi_{\mathcal{V}_r}/\Pi_{\mathcal{V}_{BP}}$), using our structure-preserving basis $\mathcal{V}_r$ and block-structured POD-basis $\mathcal{V}_{BP}$.

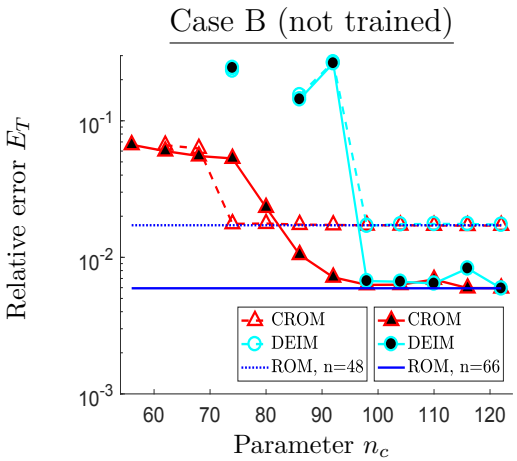


Fig. 3. Error of proposed complexity reduction CROM and non-structure-preserving DEIM for varying parameter n_c used as complexity reduction dimension. Underlying are ROMs of dimension $n = 48$ *(dashed-dotted lines)* and $n = 66$ *(solid lines)*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Angle-based formation control for a class of underlying triangulated Laman graphs

Ningbo Li, * **Pablo Borja,** ** **Arjan van der Schaft,** *
Jacquelien M. A. Scherpen, *

* *Jan C. Willems Center for Systems and Control, University of Groningen, The Netherlands (e-mail: ningbo.li@rug.nl, a.j.van.der.schaft@rug.nl, and j.m.a.scherpen@rug.nl).*
** *Department of Cognitive Robotics, Delft University of Technology, Delft, The Netherlands (e-mail: l.p.borjarosales@tudelft.nl).*

Abstract: This extended abstract proposes a passivity-based approach using bearing and velocity information for an angle-based formation control with a class of underlying triangulated Laman graphs. The controller is designed using virtual couplings on the relative measurements related to the edges. The different embedding of the graph is mapped by the measurement Jacobian, which is calculated by the time-evolution of the measurement. Furthermore, to avoid unavailable distance measurements in the control law, an estimator is designed based on the port-Hamiltonian theory using bearing and velocity measurements. The stability analysis of the closed-loop system is provided and simulations are performed to illustrate the effectiveness of the approach.

1. INTRODUCTION

Over the last three decades, formation control has attracted extensive interest due to its potential applications in many domains. Recently, the passivity-based port-Hamiltonian (pH) approach has been used for the design of formation controllers, such as Vos et al. (2014), Stacey and Mahony (2015), Xu and Liang (2018). This approach not only allows for complex and heterogenous agent dynamics but also enables the flexibility and scalability of the network.

In terms of the sensing capability, using partial information of the positions of agents requires fewer onboard sensors, which reduces the cost of hardware and introduces fewer measurement errors. Much research has been reported on this topic in recent years, such as Anderson et al. (2008), Cao et al. (2011) for distance measurement, Zhao et al. (2019), Trinh et al. (2018) for bearing measurement, and Chen et al. (2020), Jing et al. (2019) for angle measurement. In this extended abstract, we study the case where the sensing capability of the agents is based on bearing and relative velocity measurements, and angles constrain the interaction topology of agents. Remarkably, angle-based constraints are expressed by less information of the agents compared with position-, distance- and bearing-based approaches. Therefore, it is invariant to more group motions, such as translation, rotation, scaling and reflection, which means the agents can perform these maneuvers while satisfying angle-based constraints.

We consider a particular class of undirected, triangulated Laman graphs $\mathcal{G}_N(\mathcal{V}_N, \mathcal{E})$ introduced in Chen et al. (2017), where the interaction topology is determined only by angles. According to Jing et al. (2019), if a triangulated Laman framework $(\mathcal{G}_N, q)$ is strongly nondegenerate, then it is globally angle rigid. Therefore, the only realization of the framework with the underlying graph $\mathcal{G}_N$ is guaranteed. Since the topology of any formation shape constrained by angles can be designed as an underlying triangulated Laman graph, our proposed controller is not restrictive on formation shape.

In this work, we adopt a passivity-based approach, where the control objectives are achieved by virtual couplings where the virtual springs determine the formation by shaping the potential energy function of the network, while the virtual dampers shape the transient response by injecting damping. Customarily, controllers resulting from passivity-based approaches require the agents to have complete information of the relative positions even if the sensing capability and the interaction topology of the agents are both only angles. To solve this problem, we extend the passive adaptive compensator proposed for bearing formation control in Stacey and Mahony (2015) to estimate the unavailable distance information by using relative velocity.

Examples of recent literature on the angle-based formation problem are Basiri et al. (2010) and Chen et al. (2020), where an intuitive control law is proposed using only local bearing information while proving stability via linearization. However, a suitable Lyapunov function is not given for stability analysis and only single integrator models are considered. In Jing et al. (2019), a gradient-like control law is proposed for a single integrator model using bearing and distance information. Although only bearing and distance measurement are used to achieve the formation stabilization, when considering the trajectory tracking for complex dynamics, the velocity measurement is also needed. In contrast to the mentioned reference, our approach can achieve the trajectory tracking for complex dynamics without distance measurement.

The contributions of our approach are summarized as follows:

(i) Existing research only considers the single integrator case. In this work, we propose a control law and and estimator based on virtual couplings and pH theory for a double integrator model. Due to the pH framework, this approach is not only applicable to angle information but also for other measurements, such as displacement, bearing, and

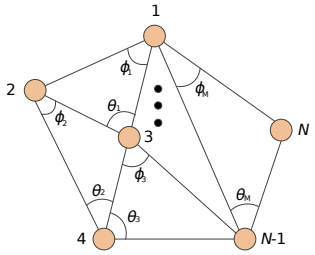


Fig. 1. Triangulated Laman graph with M triangles

distance. Therefore, our approach gives a general framework for the heterogeneous information of a multi-agent system.

(ii) We use less information to achieve more formation maneuvers. We not only consider the formation stabilization but also several maneuvers, such as velocity tracking, scale and orientation control. In particular, compared with displacement-based, distance-based, bearing-based formation, the angle-based formation can achieve more maneuvers while satisfying the constraints. Moreover, since a pH-based distance estimator is designed using bearing and velocity measurements, our approach achieves these maneuvers without using distance measurement.

The rest of the abstract is structured as follows: the problem formulation is introduced in Section 2, the control architecture is developed in Section 3, and the formation maneuvering design is provided in Section 4.

2. PROBLEM FORMULATION

Consider a group of N agents. The dynamics of the agents are given by double integrators on $\mathbb{R}^2$, which are expressed in pH form as

$$\begin{pmatrix}\dot{q}_n\\ \dot{p}_n\end{pmatrix} = \begin{pmatrix}0 & I_2\\ -I_2 & 0\end{pmatrix}\begin{pmatrix}\frac{\partial H_n}{\partial q_n}(p_n)\\ \frac{\partial H_n}{\partial p_n}(p_n)\end{pmatrix} + \begin{pmatrix}0\\ I_2\end{pmatrix}U_n, \tag{1}$$

$$H_n(p_n) = \frac{1}{2m_n}p_n^T p_n,\ \ Y_n = \frac{\partial H_n}{\partial p_n}(p_n),\ \ n \in \{1,2,...,N\}$$

where $q_n = (q_{x_n}, q_{y_n}) \in \mathbb{R}^2$, $p_n = (p_{x_n}, p_{y_n}) \in \mathbb{R}^2$, and m_n denote the position, momentum, and mass of agent n, respectively; $U_n = (U_{x_n}, U_{y_n}) \in \mathbb{R}^2$ and $Y_n = (Y_{x_n}, Y_{y_n}) \in \mathbb{R}^2$ denote the input and output, respectively; and $H_n : \mathbb{R}^2 \to \mathbb{R}$ represents the Hamiltonian of agent n.

We assume that each agent has access to bearing and relative velocity information. For relative velocity information, each agent is either able to estimate and communicate its own velocity or measure the relative velocity directly. We consider that the group of agents is connected by an underlying triangulated Laman graph $\mathcal{G}_N(\mathcal{V}_N, \mathcal{E})$ with M triangles as shown in Fig.1. We refer the reader to Chen et al. (2017) for specific definitions. Note that the graph considered in this work is undirected and the angle rigidity is ensured by Lemma 1, Jing et al. (2019).

Lemma 1. A triangulated Laman framework $(\mathcal{G}_N, q)$ is strongly nondegenerate, i.e., the three edges of each triangle are not collinear, only if $(\mathcal{G}_N, q)$ is globally angle rigid.

Note that the framework is composed of the graph $\mathcal{G}_N$ and the realization q, where *angle rigid* means that the formation shape is uniquely determined up to translations, scalings, and rotations. For more details about angle rigid, see Chen et al. (2020) and Jing et al. (2019). In particular, a triangulated Laman graph is constructed from a line graph with two nodes, every new adding node is connected by two existing nodes which are also connected. As shown in Fig. 1, the formation shape is determined by the marked angles. Since every triangulated Laman graph can be decomposed into several triangular graphs, we first design the controller for a specific triangular graph and then extend it to a general triangulated Laman graph.

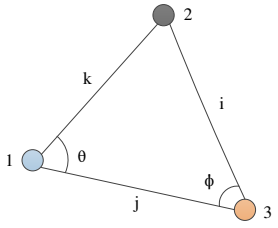


Fig. 2. Triangular formation

3. FORMATION STABILIZATION

3.1 Controller for triangular formation

Consider the triangular case shown in Fig.2, where $1,2,3$ are agents and i,j,k are the the edges. Angle 1 and 2 are the angles to be controlled. Then, the relative position, distance, and bearing of the edge k are defined as

$$z_k = q_1 - q_2, \quad r_k = ||z_k||, \quad s_k = \frac{z_k}{||z_k||},$$

respectively. Since the cosine function is monotone in the domain of the inner angle, i.e., $[0,\pi]$, we use the cosine of the angle to represent the angle measurement, which can be calculated via bearing measurement. In particular, for the angles θ and ϕ, we have

$$\cos\theta = s_k^T s_j, \quad \cos\phi = s_i^T s_j.$$

Now we consider the controller of agent 1. Since the moving of agent 1 affects both θ and ϕ, the controller of the agent 1 consist of two parts. One is to satisfy the constraint of the angle θ, the other is to satisfy the constraint of the angle ϕ. The control aim is to design a controller to ensure the cosine of θ , given by $(s_k^T s_j)$, to converge to the desired value $(s_k^T s_j)^*$. Hence, we define the error as

$$\widetilde{(s_k^T s_j)} := (s_k^T s_j) - (s_k^T s_j)^*. \tag{2}$$

It is necessary to assign a new potential energy related to the angle to ensure that the system converges to the desired point, i.e., the error converges to zero. To this end, we propose the following Hamiltonian function

$$H_{\theta 1} = \frac{1}{2}c_{\theta 1}\widetilde{(s_k^T s_j)}^2, \tag{3}$$

where $c_{\theta 1} > 0$ is a constant. Hence, the corresponding controller with spring and damping term can be derived as

$$\dot{\widetilde{(s_k^T s_j)}} = \omega_{\theta 1}, \quad \gamma_{\theta 1} = \frac{\partial H_{\theta 1}}{\partial \widetilde{(s_k^T s_j)}} + d_{\theta 1}\omega_{\theta 1}, \tag{4}$$

where $\omega_{\theta 1}$ denotes the input of the controller, $d_{\theta 1} > 0$ is a positive constant. Note that $\gamma_{\theta 1}$ is the resulting virtual force in the space of angle measurements. According to the pH theory van der Schaft and Jeltsema (2014), we define the force and

velocity as effort and flow, respectively. Thus, the power of the port can be derived as

$$< \gamma_{\theta 1} | \frac{\mathrm{d}(s_k^T s_j)}{\mathrm{d}t} >= \gamma_{\theta 1}^T \frac{\mathrm{d}(s_k^T s_j)}{\mathrm{d}t}. \tag{5}$$

Note that we only consider the relation between the agent 1 and the angle θ. To transform the power from angle measurement space to $\mathbb{R}^2$, we compute

$$\begin{aligned} < \gamma_{\theta 1} | \frac{\mathrm{d}(s_k^T s_j)}{\mathrm{d}t} > &=< \gamma_{\theta 1} | L_{\theta 1} \dot{q}_1 > \\ &=< -L_{s_k}^T s_j \gamma_{\theta 1} | \dot{q}_1 > + < -L_{s_j}^T s_k \gamma_{\theta 1} | \dot{q}_1 >, \end{aligned} \tag{6}$$

where $L_{s_k} = \frac{1}{r_k}(I_2 - s_k s_k{}^T) \in \mathbb{R}^{2\times 2}$ and $L_{s_j} = \frac{1}{r_j}(I_2 - s_j s_j{}^T) \in \mathbb{R}^{2\times 2}$. The effort of the port in (6) relies on the distance information which is not measurable. In order to avoid distance measurement, we use the relative velocity measurement to estimate the unknown distance Duindam et al. (2009).

Note that the estimated distance is used, the distance term in the angle Jacobian also needs to be replaced, accordingly. Therefore, the estimated angle Jacobian is given by

$$\begin{aligned} \hat{L}_{\theta 1} &= s_j^T \hat{L}_{\theta s_k} + s_k^T \hat{L}_{\theta s_j} \\ &= s_j^T \frac{1}{\hat{r}_{\theta k}}(I_2 - s_k s_k{}^T) + s_k^T \frac{1}{\hat{r}_{\theta j}}(I_2 - s_j s_j{}^T), \end{aligned} \tag{7}$$

where $\hat{r}_{\theta k}$ is the estimate of the edge k using the measurement of the angle θ, and $\hat{r}_{\theta j}$ is the estimate of the edge j using the measurement of the angle θ. Correspondingly, $\hat{L}_{\theta s_k}, \hat{L}_{\theta s_j}$ are the estimated bearing Jacobians using the measurement of θ. However, if $\hat{L}_{\theta 1}$ is used to replace $L_{\theta 1}$ in the right-hand side of (6), the equation is not satisfied because the effort $\gamma_{\theta 1}$ corresponds to the real flow (i.e., the time derivative of $s_k^T s_j$) in the angle space. This causes a discrepancy in the power through the virtual coupling due to the error between the estimated distance and the real unknown distance.

Define the distance errors

$$\bar{r}_{\theta k} := \hat{r}_{\theta k} - r_k, \quad \bar{r}_{\theta j} := \hat{r}_{\theta j} - r_j.$$

Then, the estimated effort in $\mathbb{R}^2$ is

$$-(\hat{L}_{\theta s_k}^T s_j + \hat{L}_{\theta s_j}^T s_k)\gamma_{\theta 1} = -L_{s_k}^T s_j \alpha_{\theta k} - L_{s_j}^T s_k \alpha_{\theta j}, \tag{8}$$

where $\alpha_{\theta k} := \frac{r_k}{\hat{r}_{\theta k}}\gamma_{\theta 1}$ is the estimated effort related to $\hat{r}_{\theta k}$ and $\alpha_{\theta j} := \frac{r_j}{\hat{r}_{\theta j}}\gamma_{\theta 1}$ is the estimated effort related to $\hat{r}_{\theta j}$. Furthermore, considering the ports in different spaces, we have

$$\begin{aligned} < \hat{L}_{\theta 1}^T \gamma_{\theta 1} | \dot{q}_1 > &=< -L_{s_k}^T s_j \alpha_{\theta k} - L_{s_j}^T s_k \alpha_{\theta j} | \dot{q}_1 > \\ &=< \alpha_{\theta k} | \frac{\mathrm{d}(s_k^T s_j)}{\mathrm{d}t} > + < \alpha_{\theta j} | \frac{\mathrm{d}(s_k^T s_j)}{\mathrm{d}t} > . \end{aligned} \tag{9}$$

Comparing (6) with (9), the discrepancy between the real effort and the estimated effort can be derived as

$$\begin{aligned} \beta_{\theta k} &= \alpha_{\theta k} - \gamma_{\theta 1} = -\frac{\bar{r}_k}{\hat{r}_{\theta k}}\gamma_{\theta 1}, \\ \beta_{\theta j} &= \alpha_{\theta j} - \gamma_{\theta 1} = -\frac{\bar{r}_j}{\hat{r}_{\theta j}}\gamma_{\theta 1}. \end{aligned} \tag{10}$$

Hence, the power of the ports, with $\beta_{\theta k}$, $\beta_{\theta j}$ as the efforts, are

$$< \beta_{\theta k} | - s_j^T L_{s_k} \dot{q}_1 >, \quad < \beta_{\theta j} | - s_k^T L_{s_j} \dot{q}_1 > . \tag{11}$$

To account for the power associated with the ports in distance space, we define the corresponding Hamiltonian as

$$H_{\theta k} := \frac{1}{2} c_{\theta k} \bar{r}_{\theta k}^2, \quad H_{\theta j} := \frac{1}{2} c_{\theta j} \bar{r}_{\theta j}^2, \tag{12}$$

where $c_{\theta k}, c_{\theta j} > 0$ are constants. Then, the power of the ports in distance space are given by

$$\begin{aligned} < \frac{\partial H_{\theta k}}{\partial \bar{r}_{\theta k}} | \dot{\bar{r}}_{\theta k} > &=< c_{\theta k} \bar{r}_{\theta k} | \dot{\bar{r}}_{\theta k} >, \\ < \frac{\partial H_{\theta j}}{\partial \bar{r}_{\theta j}} | \dot{\bar{r}}_{\theta j} > &=< c_{\theta j} \bar{r}_{\theta j} | \dot{\bar{r}}_{\theta j} > . \end{aligned} \tag{13}$$

Since energy is coordinate free, the power in angle space and distance space are the same. Therefore, comparing (11) and (13), we have

$$\begin{aligned} < \beta_{\theta k} | - s_j^T L_{s_k} \dot{q}_1 > &= < c_{\theta k} \bar{r}_{\theta k} | \dot{\bar{r}}_{\theta k} > \\ \Rightarrow \qquad \dot{\bar{r}}_{\theta k} &= -\frac{\gamma_{\theta 1}^T}{c_{\theta k} \hat{r}_{\theta k}}(-s_j^T L_{s_k} \dot{q}_1). \end{aligned} \tag{14}$$

Furthermore, the dynamics of the estimators are given by

$$\dot{\hat{r}}_{\theta k} = \dot{r}_k + \dot{\bar{r}}_{\theta k} = s_k^T \dot{z}_k - \frac{\gamma_{\theta 1}^T}{c_{\theta k} \hat{r}_{\theta k}}(-s_j^T L_{s_k} \dot{q}_1). \tag{15}$$

Similarly,

$$\dot{\hat{r}}_{\theta j} = \dot{r}_j + \dot{\bar{r}}_{\theta j} = s_j^T \dot{z}_j - \frac{\gamma_{\theta 1}^T}{c_{\theta j} \hat{r}_{\theta j}}(-s_k^T L_{s_j} \dot{q}_1). \tag{16}$$

Note that we only require the relative velocity and bearing measurement in the above estimators, while the information of distance measurement is not used.

The control law of the agent 1 for the angle θ is given by

$$\begin{aligned} U_{\theta 1} =& [\frac{1}{\hat{r}_{\theta k}}(I_2 - s_k s_k{}^T)^T s_j + \frac{1}{\hat{r}_{\theta j}}(I_2 - s_j s_j{}^T)^T s_k] \\ & \times [c_{\theta 1}\widetilde{(s_k^T s_j)} + d_{\theta 1}\widetilde{\dot{(s_k^T s_j)}}]. \end{aligned} \tag{17}$$

Now, we design the controller of the agent 1 to control the angle ϕ. To this end, define the corresponding Hamiltonian as

$$H_{\phi 1} := \frac{1}{2} c_{\phi 1} \widetilde{(s_i^T s_j)}^2, \tag{18}$$

where $c_{\phi 1} > 0$ is a constant. The controller with spring and damping term is given by

$$\widetilde{\dot{(s_i^T s_j)}} = \omega_{\phi 1}, \quad \gamma_{\phi 1} = \frac{\partial H_{\phi 1}}{\partial \widetilde{(s_i^T s_j)}} + d_{\phi 1}\omega_{\phi 1}, \tag{19}$$

where $\omega_{\phi 1}$ denotes the input of the controller and $d_{\phi 1} > 0$ is a constant. Considering the ports in different spaces, we have that

$$\begin{aligned} < \alpha_{\phi 1} | (-s_i^T L_{s_j} \dot{q}_1) > &=< \hat{L}_{\phi 1}^T \gamma_{\phi 1} | \dot{q}_1 >, \\ \hat{L}_{\phi 1} = s_i^T \hat{L}_{\phi s_k} &= s_i^T \frac{1}{\hat{r}_{\phi k 1}}(I_2 - s_k s_k{}^T), \end{aligned} \tag{20}$$

where $\hat{L}_{\phi 1}$ is the estimated angle Jacobian mapping from position of the agent 1 to the angle ϕ; $\hat{L}_{\phi s_k}$ is the estimated bearing Jacobian using the measurement of the angle ϕ; and $\hat{r}_{\phi k 1}$ is the estimated distance of the edge k by the agent 1 using the measurement of the angle ϕ. Furthermore, taking the same steps as for θ, we have the following estimator

$$\dot{\hat{r}}_{\phi k 1} = \dot{r}_k + \dot{\bar{r}}_{\phi k 1} = s_k^T \dot{z}_k - \frac{\gamma_{\phi 1}^T}{c_{\phi k} \hat{r}_{\phi k 1}}(-s_i^T L_{s_j} \dot{q}_1), \tag{21}$$

where $c_{\phi k} > 0$ is a constant. Correspondingly, the control law of the agent 1 for the angle ϕ is given by

$$U_{\phi 1} = \frac{1}{\hat{r}_{\phi k 1}}(I_2 - s_k s_k{}^T)^T s_i [c_{\phi 1}\widetilde{(s_i^T s_j)} + d_{\phi 1}\widetilde{\dot{(s_i^T s_j)}}]. \tag{22}$$

Since the design process for agents 2 and 3 is similar to the process for agent 1, we omit the details.

3.2 Extension to triangulated Laman graph

Assume that the topology is designed as a class of a triangulated Laman graph $\mathcal{G}_N(\mathcal{V}_N, \mathcal{E})$ with M triangles as shown in Fig.1. In each triangle, the control law is designed as in Section 3.1. To satisfy all the angle constraints, the general control law for the whole group is sum of the control laws derived from all related triangles. Therefore, the control law for each agent is the sum of all the corresponding control laws introduced by the triangles the agent forms. Moreover, the corresponding Hamiltonian and control law are given as follows

$$\begin{aligned} H =& \frac{1}{2}(m_1\dot{q}_1^T\dot{q}_1 + m_2\dot{q}_2^T\dot{q}_2 + ... + m_N\dot{q}_N^T\dot{q}_N) \\ & \frac{1}{2}(c_{\theta_1}\widetilde{\cos\theta_1}^2 + c_{\phi_1}\widetilde{\cos\phi_1}^2 + c_{\theta_2}\widetilde{\cos\theta_2}^2 + c_{\phi_2}\widetilde{\cos\phi_2}^2 \\ & + ... + c_{\theta_M}\widetilde{\cos\theta_M}^2 + c_{\phi_M}\widetilde{\cos\phi_M}^2) \\ & + \frac{1}{2}\sum_{\varepsilon\in\mathcal{E}} c_\varepsilon \bar{r}_\varepsilon^2, \end{aligned} \tag{23}$$

$$\begin{aligned} U =& \sum_{m=1}^{M}\sum_{n\in\mathcal{N}_m}(U_{\theta_m n} + U_{\phi_m n}) \\ =& \sum_{m=1}^{M}\sum_{n\in\mathcal{N}_m}(\hat{L}_{\theta_m n}(c_{\theta_m n}\widetilde{\cos\theta_n} + d_{\theta_m n}\frac{\mathrm{d}(\cos\theta_\mathrm{m})}{\mathrm{d}t}) \\ & + \hat{L}_{\phi_m n}(c_{\phi_m n}\widetilde{\cos\phi_m} + d_{\phi_m n}\frac{\mathrm{d}(\cos\phi_\mathrm{m})}{\mathrm{d}t})), \end{aligned} \tag{24}$$

where $\mathcal{N}_m$ is the index set of agents forming the triangle m. To ensure the stability of the closed-loop system, we first give the following conjecture.

Conjecture 1. If any three agents $n \in \mathcal{N}_m$ forming the triangle m are neither coincident nor collinear at $t_0 \in \mathbb{R}_{\geq 0}$. Then, the matrix $[\hat{L}_{\theta_m n} \quad \hat{L}_{\phi_m n}]$ for the triangle m is full column rank for any $t \geq t_0$.

The main result for formation stabilization is given by the following theorem.

Theorem 1. Consider a triangulated Laman graph $\mathcal{G}_n(\mathcal{V}_n, \mathcal{E})$ with with m triangles. Under Conjecture 1 and the proposed control law (24), the group of agents converges to the desired formation constrained by the angles.

4. FORMATION MANEUVERING

To complete a certain task, such as a group of agents move an object together along a desired trajectory, formation stabilization is not enough, it is necessary to maneuver the whole formation. In this regard, we design controllers for scale and direction control, and velocity tracking, respectively.

For scale and orientation control, without loss of generality, we choose the edge k, connecting the reference agents 1 and 2, as the reference edge. Assume the coordinates of the mentioned agents are aligned and the desired displacement with prespecified scale and orientation can be described as $z_k^* = q_1^* - q_2^*$. Then, the corresponding Hamiltonian is given by

$$H_k = \frac{1}{2}||z_k - z_k^*||^2. \tag{25}$$

Then, the control law is

$$u_1 = -\frac{\partial H_k}{\partial q_1} - d_z\dot{z}_k, \quad u_2 = -\frac{\partial H_k}{\partial q_2} + d_z\dot{z}_k, \tag{26}$$

where $d_z \in \mathbb{R}^{+2\times 1}$ is a constant matrix.

For velocity tracking, we use the leader-follower strategy and consider reference agents as the leaders, which know the desired velocity v^*. Hence, the corresponding Hamiltonian is

$$H_i^v = \frac{1}{2m_i}(p_i - p^*)^T(p_i - p^*) - p_i^T v^*,$$

where p^* is the desired momentum. Furthermore, the control law is

$$U_i = -d_r v^* - d_v(v_i - v^*), \tag{27}$$

where $d_v \geq \varepsilon I_2 > 0$.

The main result of this work is given by the following theorem.

Theorem 2. Consider a group of agents modeled as in (1) and connected by a triangulated Laman graph $\mathcal{G}_n(\mathcal{V}_n, \mathcal{E})$ with M triangles. Under Conjecture 1, the desired formation shape is achieved by the control law proposed in (24), while the desired scale and orientation are achieved by the control law (26), and the velocity tracking is achieved by the control law (27).

REFERENCES

Anderson, B.D., Yu, C., Fidan, B., and Hendrickx, J.M. (2008). Rigid graph control architectures for autonomous formations. *IEEE Control Systems Magazine*, 28(6), 48–63.

Basiri, M., Bishop, A.N., and Jensfelt, P. (2010). Distributed control of triangular formations with angle-only constraints. *Systems & Control Letters*, 59(2), 147–154.

Cao, M., Yu, C., and Anderson, B.D. (2011). Formation control using range-only measurements. *Automatica*, 47(4), 776–781.

Chen, L., Cao, M., and Li, C. (2020). Angle rigidity and its usage to stabilize multi-agent formations in 2d. *IEEE Transactions on Automatic Control*.

Chen, X., Belabbas, M.A., and Başar, T. (2017). Global stabilization of triangulated formations. *SIAM Journal on Control and Optimization*, 55(1), 172–199.

Duindam, V., Macchelli, A., Stramigioli, S., and Bruyninckx, H. (2009). *Modeling and control of complex physical systems: the port-Hamiltonian approach*. Springer Science & Business Media.

Jing, G., Zhang, G., Lee, H.W.J., and Wang, L. (2019). Angle-based shape determination theory of planar graphs with application to formation stabilization. *Automatica*, 105, 117–129.

Stacey, G. and Mahony, R. (2015). A passivity-based approach to formation control using partial measurements of relative position. *IEEE Transactions on Automatic Control*, 61(2), 538–543.

Trinh, M.H., Zhao, S., Sun, Z., Zelazo, D., Anderson, B.D., and Ahn, H.S. (2018). Bearing-based formation control of a group of agents with leader-first follower structure. *IEEE Transactions on Automatic Control*, 64(2), 598–613.

van der Schaft, A.J. and Jeltsema, D. (2014). Port-Hamiltonian systems theory: An introductory overview. *Foundations and Trends® in Systems and Control*, 1(2-3), 173–378.

Vos, E., Scherpen, J., van der Schaft, A., and Postma, A. (2014). Formation control of wheeled robots in the port-hamiltonian framework. *IFAC Proceedings Volumes*, 47(3), 6662–6667.

Xu, M. and Liang, Y. (2018). Formation flying on elliptic orbits by hamiltonian structure-preserving control. *Journal of Guidance, Control, and Dynamics*, 41(1), 294–300.

Zhao, S., Li, Z., and Ding, Z. (2019). Bearing-only formation tracking control of multiagent systems. *IEEE Transactions on Automatic Control*, 64(11), 4541–4554.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Exhaustive synthesis of microwave circuits with frequency dependent couplings

Martine Olivi * **Fabien Seyfert** * **Ke-Li Wu** ** **Yan Zhang** **

* *INRIA,Factas Team, Sophia-Antipolis,France,martine.olivi@inria.fr,fabien.seyfert@inria.fr*
** *Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, klwu@cuhk.edu.hk, yzhang@link.cuhk.edu.hk*

Abstract: The synthesis of bandpass microwave filters is based on the use of equivalent circuit models made of coupled resonators. These couplings are usually supposed to be independent of the frequency. We present in this paper a circuit model including possibly frequency varying couplings. After presenting some of its properties we consider the associated synthesis problem and show how techniques such as Groebner basis computation and Schur analysis based extraction techniques can be used to solve the latter exhaustively.

Keywords: Filter synthesis, Circuit synthesis, Structured realization, Microwave filters

1. INTRODUCTION

Microwave filters are usually synthesized using equivalent circuit models. For band-pass filters circuits made of electromagnetically coupled resonators are considered. The synthesis of such circuits when starting from a prescribed scattering response, has been studied extensively Cameron (1999); Cameron et al. (2007b, 2002); Amari (2000). In particular, the relation between the coupling topology of the circuit, that is the way each circuit is coupled to the others, and its associated class of realizable transfer functions is now well understood Amari (1999); Seyfert and Bila (2007). It was shown that for a given coupling topology several circuits with different circuital values, might realize the same response. When the considered coupling topology has the so-called non-redundant property the set of equivalent circuits is finite and approaches based on the use of Gröbner basis Cameron et al. (2005, 2007a) and continuation techniques have been developed to solve this structured realisation problem exhaustively (see software Seyfert (2005)).

In waveguide filters, the electromagnetic couplings between resonators that are realised via irises or coupling windows are usually supposed frequency independent and modelled by an idealised electrical component called inverter. When the relative functioning frequency band of the filter is increased, or the thickness of the coupling irises widened, the frequency independence of the coupling elements no longer holds and might lead to substantial modelling errors. When the frequency dependency is modelled as linear, it was recently advocated that the coupling slope, when controllable, could serve as an extra design parameter Amari et al. (2010); He et al. (2019); Szydlowski et al. (2013).

In this work we will detail the underlying electrical synthesis problem and characterize, for a given coupling topology including certain frequency dependent elements, the set of scattering responses that can be achieved. Eventually the problem of determining all electrical circuits with a given coupling structure that realise a specified response will be shown equivalent to a multivariate algebraic problem. When the latter is zero dimensional we will show how a combination of Schur analysis based methods and Groebner basis techniques from computer algebra allows for an effective and exhaustive solution of the electrical synthesis problem.

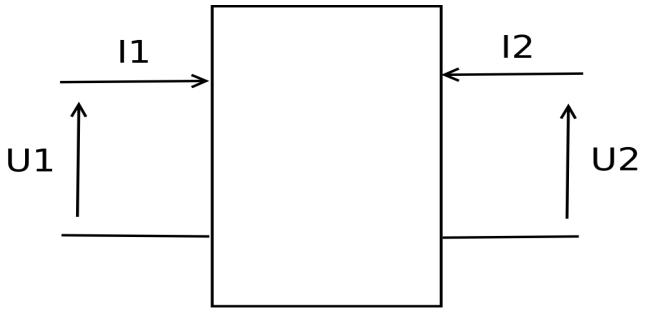


Fig. 1. Coupling structure: the inverter

2. ELECTRICAL MODEL

2.1 Frequency dependent couplings

The classical coupling element used to model small apertures in microwave filter design is the two port admittance inverter, see Figure 1. It is entirely characterised by its coupling coefficient K, and its admittance matrix is defined by following input output relation between voltages and currents at its ports:

$$\begin{cases} I_1 = jKU_2 \\ I_2 = jKU_1 \end{cases} \tag{1}$$

where we use the physicist's notation for the complex number $j^2 = -1$. The inverter is an idealised coupling element, which is introduced to model the effect of a small

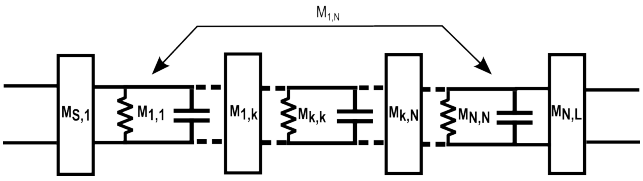


Fig. 2. Low pass equivalent circuit

aperture between coupling cavities. This complex valued electrical elements are used within low-pass equivalent circuits of filters that model, within a normalized pass-band $\omega = [-1, 1]$ the scattering behavior of the filter by means of coupled resonators, see Figure 2. If $S^h(s = j\omega)$ is the scattering matrix of the microwave filter and S^c the one of its low-pass equivalent, then

$$S^h(s) \approx S^c(as + b), \forall s \in j[-1, 1],$$

where the coefficient a, b are chosen such the frequency interval $[-1, 1]$ is mapped linearly to the actual high-frequency passband of the filter $[\omega_1, \omega_1]$. For short the low-pass equivalent circuit is a circuit with complex elements that furnishes a good approximation of a real system, i.e. the filter, within its functioning band.

As announced we introduce now a frequency variant inverter for which the admittance parameters writes,

$$\begin{cases} I_1 = j|G|\omega U_1 + j(K + G\omega)U_2 \\ I_2 = j(K + G\omega)U_1 + j|G|\omega U_2. \end{cases} \quad (2)$$

It is easily verified that the admittance matrix of the introduced frequency dependent inverter is positive real and loss-less and of McMillan degree one. The sign of the off diagonal frequency variant component G is used to model alternatively inductive or (locally) capacitiv coupling effects.

3. GENERAL LOW-PASS EQUIVALENT MODEL

We now consider low-pass circuits as represented on Fig.2, where the couplings $M_{i,j}$ between circuits can be either frequency independent, that is as described by equation (1), or frequency dependent in accordance with equation (2). The input and output couplings $M_{S,k}$ and $M_{k,L}$ are supposed frequency independent. Kirchhoff's law yields following state space equations in descriptor form for this kind of low-pass circuits,

$$\begin{cases} E\dot{V} = -jF.V + BU_{in} \\ I_{out} = B^t V \end{cases} \quad (3)$$

where E is a $n \times n$ positive definite matrix and F a symmetric matrix of same size, while B is $n \times 2$. The 2×1 voltage vector U_{in} represents the voltages at both ports of the circuit, while I_{out} represents the currents at the same locations. Eventually the state vector V is defined as $V = jU$ where $U(k)$ is the voltage in resonator k. There is of course a direct link between the electrical components of the circuit and the elements of matrix (E, F, B). If the circuit i is coupled to the circuit j in a frequency dependent manner described by the obviously defined parameters $K_{i,j}, G_{i,j}$ we have

$$E_{i,j} = E_{j,i} = G_{i,j},\ F_{i,j} = F_{j,i} = K_{i,j}.$$

If they are coupled in a frequency independent way we obviously have $E_{i,j} = 0$ and $F_{i,j} = F_{j,i} = K_{i,j}$. Eventually if they aren't coupled at all $E_{i,j} = F_{i,j} = 0$. As for the diagonal terms we have,

$$E_{i,i} = C_i + \sum_{k \neq i} |G_{k,i}|,\ F_{i,i} = M_{i,i}$$

where C_i is the capacitance of the i^{th} resonator, and $M_{i,i}$ a susceptance that allows to tune its resonant frequency. The first column of B is equal to the $M'_{S,k}s$, that is $B(k, 1) = M_{S,k}$ and the second to the $M'_{L,k}s$ that is $B(k, 2) = M_{L,k}$.

We will call circuital a realisation (E, F, B) corresponding to system (3), with E and F symmetric and E invertible.

Proposition 1. We sum up some elementary properties of these realisations. We call admittance the transfer function of a circuital realisation.

- Let (E, F, B) be a circuital realisation, E, F, B all real and E positive definite, then its admittance
$$Y = B(sE + jF)^{-1}B^t = -jB(\omega E + F)^{-1}B^t$$
is a 2×2 strictly proper, reciprocal, loss-less positive real transfer function.
- If Y is a 2×2 strictly proper, reciprocal, loss-less positive real transfer function there exists a circuital realisation (E, F, B) to it, where $E = Id$ and (F, B) are real.
- Suppose (E, F, B) and (E', F', B') are two minimal circuital realizations (possibly with complex entries) of McMillan degree n with same admittance matrix then there exists a non-singular $n \times n$ matrix P such that,
$$E' = P^t EP,\ F' = P^t FP,\ B' = P^t B.$$

4. COUPLING GRAPH AND TRANSMISSION ZEROS

The scattering matrix S of a circuit is defined as the Cayley transform of its admittance matrix,

$$S = (I - Y)(I + Y)^{-1}.$$

When Y is positive real and loss-less S is a 2×2 inner matrix. The transmission zeros are defined as the zeros of the rational function $S_{1,2}S_{2,1}$ that lie in the closed left-half plan with the convention that zeros occurring on the imaginary axis are counted with half their multiplicity (see Carlin and Civalleri (1997)). Using the Belevitch form of 2×2 inner matrices it is easily seen that their number k cannot exceed the McMillan degree n of S. If k is not maximal we say that $n - k$ transmission zeros are at infinity. The synthesis of microwave filters usually starts with the determination of an "optimal" scattering responses, passing maximally the signal in the pass-band(s) while rejecting it with a prescribed rejection level in the stop bands. The computation of such responses is done by solving quasi-convex Zolotarev problems involving the filtering function $S_{1,2}/S_{1,1}$ of the scattering matrix Lunot et al. (2008). For these problems the number of available transmission zeros is crucial, as their presence allow to obtain very selective responses with steep slopes in the transition regions between pass and stop bands. We give here a result relating the maximal number of transmission zeros a circuit can achieve to the network structure of its coupling scheme.

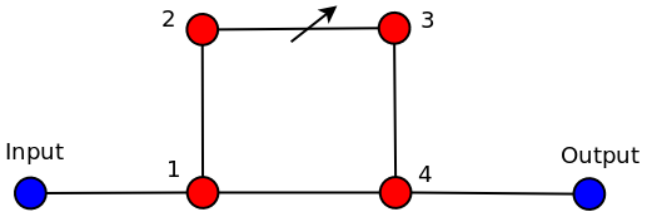


Fig. 3. Coupling graph corresponding to a four resonators filter with a "quadruplet" coupling structure. The frequency varying coupling is noted by an extra diagonal ascending arrow between resonator 2 and 3.

Proposition 2. **Shortest path rule**: To every coupled resonator circuit, we associate a coupling graph defined as described: every resonator is represented by a node, and two additional separate nodes are drawn to symbolize the input and output. Edges of length one are drawn between the input (resp. output) node and a resonator k if the corresponding $B(k,1)$ (resp. $B(k,2)$) coupling elements is present in the circuit. Edges of length one are drawn between resonators nodes k and l if a non frequency dependent coupling $F(k,l)$ (and $E(k,l)=0$) is available. Eventually an edge of length zero is drawn between resonator k and l if a frequency dependent couplings $E(k,l)$ (and $F(k,l)\neq 0$) is available in the circuit. Consider a circuit with n resonators. Let l be the length of the shortest path in the coupling graph between input and output, then the scattering matrix of this circuit can maximally possess $n+1-l$ transmission zeros.

Figure (3) represents such a coupling graph for a quadruplet structure. The shortest path between input and output is here 3, which indicates by the preceding proposition that this coupling structure can accommodate maximally $4+1-3=2$ transmission zeros. If an additional frequency varying coupling were to be placed between resonator 1 and 2 the shortest path would have a length of 2 indicating that maximally 3 transmission zeros can be produced by such a structure. Eventually note that quadruplet with independently adjustable transmission zeros are usually by means of a frequency independent cross-coupling $F(1,3)$ which is complicate to realize in practice and necessitate the placement of a wire probe between resonator 1 and 3. The use of a frequency dependent coupling that can be realized with thickened coupling irises appears here to offer an interesting alternative option for the hardware implementation of the filter.

5. CANONICAL FORM

We now come to some canonical realization of these circuits.

Proposition 3. Suppose Y is a strictly proper admittance matrix, reciprocal, positive real and loss-less of McMillan degree n. Suppose in addition that $Y_{1,1}$ and $Y_{2,2}$ are different from zeros. Let

$$Y=\sum_{k=1}^{\infty}\frac{G_k}{s^k}$$

be the formal development of Y at infinity where the $G_k's$ are the Markov parameters. The matrix Y admits a real valued circuital realization (Id,F,B) where,

- B has only three non vanishing element $B(1,1)$, $B(n,1)$, $B(n,2)$.
- The only possibly non-vanishing elements in F are its diagonal, sub and sur-diagonal, as well as it last line and last column
- If the first Markov parameter is diagonal, i.e the anti-diagonal terms of G_1 are zero, then $B(n,1)=0$. If further k first Markov parameters are diagonal $(1<k\leq n-1)$ then $F(1\dots k-1,n)=F(n,1\dots k-1)=0$
- Let S be the scattering matrix associated to Y then if S has k finite transmission zeros then the first $n-k-1$ Markov parameters of Y are diagonal.

6. SYNTHESIS OF STRUCTURED REALIZATIONS

In order to tackle the problem of synthesizing a scattering matrix with a circuit with a prescribed coupling topology we first give an algebraic meaning to the word coupling topology. A topology σ is a set of formal structured matrices (E,F,B) populated with their specific non zero elements. The parameter set X of a coupling topology is the set of formal variables representing the variable entries of E,F,B. For example the parameter set X associated to the coupling topology shown on Figure (3) is

$$X=\{B(1,1),B(2,4),F(1,1),F(2,2),F(3,3), \quad (4)$$
$$F(4,4),F(1,2),F(2,3),F(3,4),F(1,4),E(2,3)\} \quad (5)$$

while the diagonal of E is here considered as equal to the identity matrix which corresponds to a classical normalisation.

We suppose that formally (as a polynomial in $\mathbb{C}[X]$) $\det(E)\neq 0$, that is the set of singular $E's$ form a strict sub-variety $\mathcal{W}$ of $\mathbb{C}^r$, where r is the cardinality of X. To every coupling topology σ we will associate a realization map,

$$\begin{aligned}\pi_\sigma:\ \mathbb{C}^r\setminus\ \mathcal{W} &\mapsto \left(\mathbb{C}^{2\times 2}\right)^{2n-1}\\ x &\mapsto \left(B^t(x)E^{-1}(x)B(x),\, B^t(x)E^{-1}(x)F(x)E^{-1}B(x)\right. \quad (6)\\ &\quad \left.\dots\, B^t(x)(E^{-1}F)^{2n-1}(x)E^{-1}(x)B(x)\right).\end{aligned}$$

Adapting the algebraic framework detailed in Seyfert (2019) we obtain following results.

Definition 1. For a topology $\sigma=(E,F,B,X)$ with $r=card(X)$ we define $\mathcal{V}(\sigma)=\overline{\pi_\sigma(\mathbb{C}^r)}$ to be its admissible set. It is equivalent, up to the closure operation, to the set of all possible admittances the topology can generate, when its parameters range over $\mathbb{C}^r\setminus\mathcal{W}$.

Definition 2. Let $\sigma=(B,E,F,X)$ be a topology. As a polynomial map, π_σ has a Jacobian matrix. On an non-empty open Zariski set of $\mathbb{C}[X]$ this Jacobian has constant rank, which is often called its generic rank. We will say that a topology is non-redundant, if the Jacobian of its realisation map is generically full rank, that is of rank $r=card(X)$.

Solving our coupling matrix synthesis problem is about inverting π_σ on $\mathcal{V}(\sigma)$. We have following properties,

Proposition 4. Let $\sigma=(B,E,F,X)$ be a coupling topology, with $r=card(X)$. We have,

- $\mathcal{V}(\sigma)$ is an irreducible algebraic variety.
- If σ is non-redundant, then the dimension of $\mathcal{V}(\sigma)$ as an algebraic variety is r.

- If σ is non-redundant there exists an integer $\Theta(\sigma)$ such that all fibers of π_σ are generically of cardinality $\Theta(\sigma)$. More precisely there exists a non-empty Zariski set $\mathcal{U}$ open in $\mathcal{V}(\sigma)$ such that $\forall y \in U$, $card(\pi_\sigma^{-1}(y)) = \Theta(\sigma)$. U is dense in $\mathcal{V}(\sigma)$ in both topologies (Zariski and euclidean). We call $\Theta(\sigma)$ the order of the topology σ.
- If σ_1 and σ_2 are two non-redundant topologies with parameter sets of the same cardinality, and if $\mathcal{V}(\sigma_1) \subset \mathcal{V}(\sigma_2)$ then $\mathcal{V}(\sigma_1) = \mathcal{V}(\sigma_2)$
- If σ_1 and σ_2 are coupling topologies, and $\mathcal{V}(\sigma_1) \cap \mathcal{V}(\sigma_2) \neq \mathcal{V}(\sigma_1)$, then generically, on a non-empty Zariski open set $\mathcal{U}$ of $\mathcal{V}(\sigma_1)$, we have

$$\forall y \in \mathcal{U},\ \pi_{\sigma_2}^{-1}(y) = \emptyset.$$

In the talk we will detail how to solve a typical synthesis problem. Starting from a synthesized "optimal" frequency response S with a given number of transmission zeros, a canonical form (E_0, F_0, B_0) is computed. Considering a non-redundant coupling topology σ with same admissible set as the computed canonical form (specialized to the particular number of considered transmission zeros) we set up an algebraic system of multivariate polynomial equations based on the item of proposition (1) in order to find all equivalent realizations with topology σ similar (with same transfer function) to (E_0, F_0, B_0). The system is solved using Gröbner basis. For large systems we will show that for particular coupling topologies a divide and conquer strategy can be designed in order to split the original structured realisation problem in several smaller ones. This strategy uses Schur analysis to decompose the function to be realised as a sequence of chained responses of lower degree than the initial one. Details about these procedures can be found in Zhang et al. (2021a,b).

REFERENCES

Amari, S. (1999). On the maximum number of finite transmission zeros of coupled resonator filters with a given topology. *IEEE Microwave and Guided Wave Letters*, 9(9), 354–356. doi:10.1109/75.790472.

Amari, S. (2000). Synthesis of cross-coupled resonator filters using an analytical gradient-based optimization technique. *IEEE Transactions on Microwave Theory and Techniques*, 48(9), 1559–1564. doi:10.1109/22.869008.

Amari, S., Seyfert, F., and Bekheit, M. (2010). Theory of Coupled Resonator Microwave Bandpass Filters of Arbitrary Bandwidth. *IEEE Transactions on Microwave Theory and Techniques*, 58(8), 2188–2203. doi:10.1109/TMTT.2010.2052874. URL `https://hal.inria.fr/hal-00663513`.

Cameron, R.J., Harish, A.R., and Radcliffe, C.J. (2002). Synthesis of advanced microwave filters without diagonal cross-couplings. *IEEE Transactions on Microwave Theory and Techniques*, 50(12), 2862–2872. doi:10.1109/TMTT.2002.805141.

Cameron, R.J., Faugère, J.C., Rouillier, F., and Seyfert, F. (2007a). Exhaustive approach to the coupling matrix synthesis problem and application to the design of high degree asymmetric filters. *International Journal of RF and Microwave Computer-Aided Engineering*, 17(1), 4–12. doi:10.1002/mmce.20190. URL `https://hal.inria.fr/hal-00663777`.

Cameron, R.J., Faugère, J.C., and Seyfert, F. (2005). Coupling matrix synthesis for a new class of microwave filter configuration. In *2005 IEEE MTT-S International Microwave Symposium*, volume 1, 119–124. Long Beach, United States. doi:10.1109/MWSYM.2005.1516536. URL `https://hal.inria.fr/hal-00663550`.

Cameron, R. (1999). General coupling matrix synthesis methods for chebyshev filtering functions. *IEEE Transaction on Microwave Theory and Techniques*, 47(4), 433–442.

Cameron, R., Mansour, R., and Kudsia, C. (2007b). *Microwave Filters for Communication Systems: Fundamentals, Design and Applications*. Wiley. URL `https://books.google.fr/books?id=GyVTAAAAMAAJ`.

Carlin, H. and Civalleri, P. (1997). *Wideband Circuit Design*. CRC Press.

He, Y., Macchiarella, G., Ma, Z., Sun, L., and Yoshikawa, N. (2019). Advanced direct synthesis approach for high selectivity in-line topology filters comprising n - 1 adjacent frequency-variant couplings. *IEEE Access*, 7, 41659–41668. doi:10.1109/ACCESS.2019.2907531.

Lunot, V., Seyfert, F., Bila, S., and Nasser, A. (2008). Certified Computation of Optimal Multiband Filtering Functions. *IEEE Transactions on Microwave Theory and Techniques*, 56(1), 105–112. doi:10.1109/TMTT.2007.912234. URL `https://hal.inria.fr/hal-00663542`.

Seyfert, F. (2005). Software Dedale-HF. `https://www-sop.inria.fr/apics/Dedale/WebPages`.

Seyfert, F. (2019). *Méthodes analytiques pour la conception et le réglage de dispositifs micro-ondes*. Habilitation à diriger des recherches, UCA ; Université Côte d'Azur. URL `https://hal.inria.fr/tel-02444432`.

Seyfert, F. and Bila, S. (2007). General synthesis techniques for coupled resonator networks. *IEEE Microwave Magazine*, 8(5), 98–104. doi:10.1109/MMW.2007.4383440. URL `https://hal.inria.fr/hal-00663533`.

Szydlowski, L., Leszczynska, N., and Mrozowski, M. (2013). Generalized chebyshev bandpass filters with frequency-dependent couplings based on stubs. *IEEE Transactions on Microwave Theory and Techniques*, 61(10), 3601–3612. doi:10.1109/TMTT.2013.2279777.

Zhang, Y., Seyfert, F., Amari, S., Olivi, M., and Wu, K.L. (2021a). General synthesis method for dispersively coupled resonator filters with cascaded topologies. *IEEE Transactions on Microwave Theory and Techniques*, 69(2), 1378–1393. doi:10.1109/TMTT.2020.3041223.

Zhang, Y., Seyfert, F., and Wu, K.L. (2021b). Exhaustive synthesis and realization of extended-box topologies with dispersive couplings. In *2021 IEEE MTT-S International Microwave Filter Workshop (IMFW)*, 33–35. doi:10.1109/IMFW49589.2021.9642326.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A novel constraint tightening approach for robust data-driven predictive control ⋆

Christian Klöppelt * **Julian Berberich** ** **Frank Allgöwer** **
Matthias A. Müller *

* *Institute of Automatic Control, Leibniz University Hannover, Germany, (e-mail: {kloeppelt,mueller}@irt.uni-hannover.de)*
** *Institute for Systems Theory and Automatic Control, University of Stuttgart, Germany, (e-mail: {berberich,allgower}@ist.uni-stuttgart.de)*

Abstract: We present a data-driven predictive control scheme for the stabilization of unknown LTI systems subject to process disturbances. The scheme uses Willems' lemma for the prediction of future system trajectories and can be set up using only a priori measured input-output data of the disturbed system and an upper bound on its order. The main contribution is the introduction of a novel constraint tightening, which purely based on data guarantees closed-loop constraint satisfaction and recursive feasibility, even in the presence of process disturbances. Furthermore, a pre-stabilizing controller can be integrated into the scheme which ensures applicability for unstable systems.

Keywords: Model predictive control, data-driven control, robust control

1. INTRODUCTION

Recent research has focused on the idea of establishing predictive control schemes without explicit model knowledge using only a priori measured data sequences (Yang and Li, 2015; Coulson et al., 2019, 2021; Berberich et al., 2021). To this end, the above publications employ Willems' lemma (Willems et al., 2005) for the data-based prediction of future system trajectories.

One of the major strengths of model predictive control (MPC) is its capability of including state and input constraints into the optimal control problem, and therefore, guaranteeing the satisfaction of these constraints in closed-loop operation. In the data-driven setting –based on Willems' lemma– these closed-loop guarantees were so far only achieved in Berberich et al. (2020a) by the introduction of a proper constraint tightening which can be parametrized using only measured data. However, the aforementioned publication only considers additive output measurement noise, whereas process disturbances, acting directly on the states, were barely considered in the data-driven MPC literature so far.

While process disturbances are considered in Huang et al. (2021) and Umenberger (2021), knowledge of a priori measured disturbances is assumed in both publications which could be a rather restrictive assumption depending on the application. Moreover, no closed-loop guarantees were derived in both publications. Recently, a scheme guaranteeing closed-loop stability and recursive feasibility in the presence of process disturbances was introduced in Liu et al. (2021), which, however, lacks of guarantees for closed-loop constraint satisfaction.

The predictive control scheme introduced in the following closes this gap by extending the constraint tightening from Berberich et al. (2020a) to the case with process disturbances. As a key advantage, the proposed approach allows for the inclusion of a pre-stabilizing feedback, which enables the usage of the proposed scheme even in the case of unstable systems. Furthermore, a suitable input constraint tightening is introduced in order to guarantee closed-loop input constraint satisfaction while using the pre-stabilizing controller.

In this extended abstract, after introducing the problem setup in Section 2, we present the proposed data-based MPC scheme and state its main properties (closed-loop constraint satisfaction and practical exponential stability) in Section 3. More details on the results presented, as well as the associated proofs, can be found in Klöppelt et al. (2022).

Notation: We denote the set of continuous, strictly increasing functions $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\alpha(0) = 0$ by $\mathcal{K}$ and the subset of unbounded functions in $\mathcal{K}$ as $\mathcal{K}_\infty$.

For a sequence $\{z_k\}_{k=0}^{N-1}$ we define the Hankel matrix of depth L as

$$H_L(z) = \begin{bmatrix} z_0 & z_1 & \cdots & z_{N-L} \\ z_1 & z_2 & \cdots & z_{N-L+1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{L-1} & z_L & \cdots & z_{N-1} \end{bmatrix},$$

⋆ This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant MU 3929/1-2 and AL 316/12-2 - 279734922, under Germany's Excellence Strategy - EXC 2075 - 390740016, and under grant 468094890. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Julian Berberich.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

and the stacked window from time instant a to b as

$$z_{[a,b]} = \begin{bmatrix} z_a^\top & \cdots & z_b^\top \end{bmatrix}^\top.$$

2. PROBLEM SETUP

We consider the discrete-time LTI system

$$x_{k+1} = Ax_k + Bu_k + w_k \tag{1}$$

with the state $x_k \in \mathbb{R}^n$, the input $u_k \in \mathbb{R}^m$, and the process disturbance $w_k \in \mathbb{R}^n$, where the pair (A, B) is controllable. We assume that the state, input, and disturbance are subject to constraints in the form of hypercubes, i.e., the input constraint set is given by

$$\mathbb{U} = \{u \in \mathbb{R}^m \mid \|u\|_\infty \leq u_{\max}\}, \tag{2}$$

for some $u_{\max} > 0$, and analogously for the state and disturbance constraint sets $\mathbb{X}$ and $\mathbb{W}$. The control goal is to stabilize the origin in the presence of the bounded disturbance $w_k \in \mathbb{W}$ for all $k \geq 0$, while satisfying the input and state constraints $x_k \in \mathbb{X}$ and $u_k \in \mathbb{U}$ for all $k \geq 0$. We assume that the system matrices A and B are unknown, but that the system order n, as well as the constraints $u_{\max}$, $x_{\max}$, $w_{\max}$ are known. Hence, for controller design, only the latter shall be used together with measured input and state sequences. Note that the following results also hold if only an upper bound on the system order is known.

To this end, we a priori apply a persistently exciting (PE) input sequence to the system and measure the resulting state sequence, where a PE sequence is defined as follows.

Definition 1. A sequence $\{u_k\}_{k=0}^{N-1}$, with $u_k \in \mathbb{R}^m$, is persistently exciting of order L if $\operatorname{rank}(H_L(u)) = mL$.

Using the input-state data sequences we make use of Willems' fundamental lemma. This result states that in the absence of disturbances (i.e., if $w_k = 0$ for all $k \geq 0$) all system trajectories can be parametrized by the linear combination of time shifts of a priori measured system trajectories.

Lemma 1. (Willems et al. (2005)). Suppose the data sequence $\{u_k, \hat{x}_k\}_{k=0}^{N-1}$ is a trajectory of the system

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k \tag{3}$$

and u is PE of order $L + n$. Then, $\{\bar{u}_k, \bar{x}_k\}_{k=0}^{L-1}$ is a trajectory of system (3) if and only if there exists $\alpha \in \mathbb{R}^{N-L+1}$ such that

$$\begin{bmatrix} H_L(u) \\ H_L(\hat{x}) \end{bmatrix} \alpha = \begin{bmatrix} \bar{u} \\ \bar{x} \end{bmatrix}.$$

In the next section, we set up an MPC scheme which uses Lemma 1 for the prediction of future trajectories, and moreover, guarantees closed-loop constraint satisfaction.

3. DATA-DRIVEN PREDICTIVE CONTROL SCHEME

First, we assume that we have access to a state feedback matrix K such that all eigenvalues of $A_K = A + BK$ lie strictly inside the unit disc. In case of a stable system, this feedback can be set to zero. In case of an unstable system, such a control law can be computed purely from data, e.g., following the approaches in Berberich et al. (2020b); Van Waarde et al. (2020). We use this feedback matrix to pre-stabilize the system by using the input parametrization

$$u_k = Kx_k + \nu_k \tag{4}$$

as it is common, for example, in tube-based MPC (Chisci et al., 2001).

To make use of Lemma 1 for the prediction of state sequences, we consider the input ν_k of the pre-stabilized system

$$x_{k+1} = A_K x_k + B\nu_k + w_k, \tag{5}$$

apply the PE input sequence $\{\nu_k^d\}_{k=0}^{N-1}$ of length N to system (5), and measure the resulting state sequence $\{x_k^d\}_{k=0}^{N}$, where the superscript "d" is used to denote the fact that this is a priori collected data.

Assumption 1. The input sequence $\{\nu_k^d\}_{k=0}^{N-1}$ is persistently exciting of order $L + n$.

Assumption 1 can be easily enforced in practice by choosing a rich enough input sequence to generate the data. It is not restrictive in the sense that the set of signals which are not PE has measure zero.

With these a priori generated data sequences,we are now in the position to set up the optimal control problem (OCP) at time t with the prediction horizon L

$$J_L^*(x_t) = \min_{\substack{\alpha(t),\sigma(t),\\ \bar{u}(t),\bar{x}(t)}} J_L\left(\bar{u}(t), \bar{x}(t), \alpha(t), \sigma(t)\right), \tag{6a}$$

$$\text{s.t.} \begin{bmatrix} \bar{u}(t) \\ \bar{x}(t) + \sigma(t) \end{bmatrix} = \begin{bmatrix} H_L(\nu^d) \\ H_{L+1}(x^d) \end{bmatrix} \alpha(t), \tag{6b}$$

$$\bar{x}_0(t) = x_t, \tag{6c}$$

$$\bar{x}_L(t) = 0, \tag{6d}$$

$$f_k^x\left(\bar{u}(t), \bar{x}(t), \alpha(t), \sigma(t)\right) \leq x_{\max}, \tag{6e}$$

$$f_k^u\left(\bar{u}(t), \bar{x}(t), \alpha(t), \sigma(t)\right) \leq u_{\max}, \tag{6f}$$

$$\forall k = 0, \ldots, L-1, \tag{6g}$$

where (6b) is used as prediction for future state sequences $\bar{x}(t)$ (compare Lemma 1), (6c) and (6d) are the initial and terminal condition of the OCP. The terminal equality constraint is used to prove exponential stability in Theorem 1, similar to standard (model-based) MPC (Rawlings et al., 2017). Further, (6e) and (6f) are suitable tightened constraints ensuring recursive feasibility and closed-loop constraint satisfaction.

We use the cost function

$$\begin{aligned} J_L\left(\bar{u}(t), \bar{x}(t), \alpha(t), \sigma(t)\right) = &\sum_{k=0}^{L-1} \left(\|\bar{u}_k(t)\|_R^2 + \|\bar{x}_k(t)\|_Q^2 \right) \\ &+ \lambda_\alpha w_{\max} \|\alpha(t)\|_2^2 \\ &+ \frac{\lambda_\sigma}{w_{\max}} \|\sigma(t)\|_2^2, \end{aligned} \tag{7}$$

with quadratic regularization on $\alpha(t)$, and $\sigma(t)$ (compare Berberich et al. (2021)). Using the slack variable σ (first introduced in Coulson et al. (2019)) is common in data-driven predictive control, in order to ensure feasibility of (6b) even in the presence of disturbances. The state and input constraint functions are given by

$$\begin{aligned} f_k^x\left(\bar{u}(t), \bar{x}(t), \alpha(t), \sigma(t)\right) = &\|\bar{x}_k(t)\|_\infty + a_{u,k} \|\bar{u}(t)\|_1 + a_{c,k} \\ &+ a_{\alpha,k} \|\alpha(t)\|_1 + a_{\sigma,k} \|\sigma_k(t)\|_\infty \end{aligned} \tag{8}$$

$$f_k^u\left(\bar{u}(t),\bar{x}(t),\alpha(t),\sigma(t)\right) = \left\|\bar{u}_k(t)\right\|_\infty + b_{u,k}\left\|\bar{u}(t)\right\|_1 + b_{c,k} + b_{\alpha,k}\left\|\alpha(t)\right\|_1 + b_{\sigma,k}\left\|\sigma_k(t)\right\|_\infty + \left\|K\bar{x}_k(t)\right\|_\infty. \tag{9}$$

In the following, we describe how the various constants appearing in (8), (9) are defined. This is done in a similar fashion as in Berberich et al. (2020a), but suitably extended in order to be able to account for process disturbances and the additional pre-stabilizing feedback in (4). In particular, a suitable input constraint tightening resulting in (6f) and (9) was not present in Berberich et al. (2020a). First, define the constants

$$c_{\alpha,k} \coloneqq \rho_{A,k}\bar{d}_{N-L} + \bar{d}_{N-L+k}, \tag{10}$$

$$c_{\sigma,k} \coloneqq \rho_{A,k} + 1, \tag{11}$$

for $k = 0,\dots,L-1$, where $\rho_{A,k} \geq \left\|A_K^k\right\|_\infty$ and $\bar{d}_k \geq \left\|d_k\right\|_\infty$, with

$$d_k \coloneqq \sum_{i=0}^{k-1} A_K^{k-1-i} w_i. \tag{12}$$

Moreover, note that $\rho_{A,k}$ and $\bar{d}_k$ can be computed based on the available input-state data and the disturbance bound $w_{\max}$, following the approach of Wildhagen et al. (2022, Section IV.B). Furthermore, we consider a controllability constant Γ, which also can be computed from data (Berberich et al., 2020a, Section V.A), and the constant $c_{pe} = \left\|H_{u\hat{x}}^\dagger\right\|_1$, with

$$H_{u\hat{x}} = \begin{bmatrix} H_L\left(\nu^d\right) \\ H_1\left(\hat{x}^d_{[0,N-L]}\right) \end{bmatrix}, \tag{13}$$

where $\hat{x}^d_{[0,N-L]}$ denotes the a priori measured state sequence without process disturbances. We approximate $c_{pe} \approx \left\|H_{ux}^\dagger\right\|_1$, where H_{ux} is analogous to (13) but contains disturbances in the measured state sequence. However, the error between the real value and the approximation of c_{pe} is small for sufficiently small disturbance levels $w_{\max}$, as was confirmed by numerical simulations.

Using these system constants, as well as $\bar{K} = \left\|K\right\|_\infty$, we can now define the coefficients of the state constraint tightening (8) as

$$a_{u,k} = 0,\ a_{\alpha,k} = c_{\alpha,k},\ a_{\sigma,k} = c_{\sigma,k},\ a_{c,k} = \bar{d}_k, \tag{14}$$

for $k = 0,\dots,n-1$, and

$$\begin{aligned} a_{u,k+n} &= a_{u,k} + a_{\alpha,k}c_{pe} + a_{\sigma,k}c_{pe}\bar{d}_n, \\ a_{\alpha,k+n} &= a_{u,k+n}\Gamma c_{\alpha,L-1} + c_{\alpha,k+n}, \\ a_{\sigma,k+n} &= a_{u,k+n}\Gamma c_{\sigma,L-1} + c_{\sigma,k+n}, \\ a_{c,k+n} &= \bar{d}_n + a_{c,k} + a_{\alpha,k}\left(nx_{\max} + n\bar{d}_n\right) \\ &\quad + a_{\sigma,k}\left(\bar{d}_{N-1}c_{pe}\left(nx_{\max} + n\bar{d}_n\right) + \bar{d}_n\right). \end{aligned} \tag{15}$$

for $k = 0,\dots,L-n-1$. Further, we define the coefficients for the input constraint tightening (9) as

$$b_{u,k} = 0, b_{\alpha,k} = \bar{K}c_{\alpha,k}, b_{\sigma,k} = \bar{K}c_{\sigma,k}, b_{c,k} = \bar{K}\bar{d}_k, \tag{16}$$

for $k = 0,\dots,n-1$, and

$$\begin{aligned} b_{u,k+n} &= b_{u,k} + b_{\alpha,k}c_{pe} + b_{\sigma,k}c_{pe}\bar{d}_n, \\ b_{\alpha,k+n} &= b_{u,k+n}\Gamma c_{\alpha,L-1} + \bar{K}c_{\alpha,k+n}, \\ b_{\sigma,k+n} &= b_{u,k+n}\Gamma c_{\sigma,L-1} + \bar{K}c_{\sigma,k+n}, \\ b_{c,k+n} &= \bar{K}\bar{d}_n b_{c,k} + b_{\alpha,k}\left(nx_{\max} + n\bar{d}_n\right) \\ &\quad + a_{\sigma,k}\left(\bar{d}_{N-1}c_{pe}\left(nx_{\max} + n\bar{d}_n\right) + \bar{d}_n\right), \end{aligned} \tag{17}$$

for $k = 0,\dots,L-n-1$. At this point it becomes clear why in case of an unstable system a pre-stabilizing controller (4) should be included into the control scheme, as in case of eigenvalues of A lying on or outside the unit disc the constants $\rho_{A,k}$ and $\bar{d}_k$ would (exponentially) diverge. Thus, the coefficients (14)-(17) would diverge as well, resulting in an infeasible OCP (6) due to (6e), (6f) even for small prediction horizons L.

The strictly convex OCP (6) is now solved in an n-step receding horizon manner, i.e., it is solved at time t with the measured state x_t, the first n optimal inputs $\bar{u}^*_{[0,n-1]}(t)$ are applied to system (1) via the input parametrization (4), i.e., $u_{t+k} = Kx_{t+k} + \bar{u}^*_k(t)$ for $k = 0,\dots,n-1$. This procedure is then repeated at time $t + jn$ for all $j \in \mathbb{N}$. Next, we state that for all states inside the region of attraction, i.e., with $J_L^*(x_t) \leq V_{\mathrm{ROA}}$, a sufficiently small disturbance bound $w_{\max}$ can be found such that the MPC scheme results in recursive feasibility, practical exponential stability, and closed-loop constraint satisfaction, i.e., $x_t \in \mathbb{X}$ and $u_t \in \mathbb{U}$ for all $t \geq 0$.

Theorem 1. Suppose that Assumption 1 holds. Then, for any $V_{\mathrm{ROA}} > 0$, there exist $\underline{\lambda}_\alpha$, $\overline{\lambda}_\alpha$, $\underline{\lambda}_\sigma$, $\overline{\lambda}_\sigma > 0$ such that for all λ_α, λ_σ satisfying

$$\underline{\lambda}_\alpha \leq \lambda_\alpha \leq \overline{\lambda}_\alpha, \quad \underline{\lambda}_\sigma \leq \lambda_\sigma \leq \overline{\lambda}_\sigma, \tag{18}$$

there exist $\bar{w}$, $\bar{c}_{pe} > 0$ as well as a function $\beta \in \mathcal{K}_\infty$, such that for all $w_{\max}$ and c_{pe} satisfying

$$w_{\max} \leq \min\left\{\bar{w},\ \frac{\bar{c}_{pe}}{c_{pe}}\right\}, \tag{19}$$

the following holds for the closed-loop of the n-step MPC scheme:

(i) If $J_L^*(x_t) \leq V_{\mathrm{ROA}}$ for some $t \geq 0$, then OCP (6) is feasible at time $t + n$.
(ii) For any initial condition satisfying $J_T^*(x_0) \leq V_{\mathrm{ROA}}$ it holds that $x_t \in \mathbb{X}$ and $u_t \in \mathbb{U}$ for all $t \geq 0$, and $J_L^*(x_t)$ converges exponentially to $J_L^*(x_t) \leq \beta(\bar{w})$.

Note that (ii) only shows exponential convergence of x_t to a neighborhood of $x = 0$, however, it is possible to establish a suitable lower as well as an upper bound on $J_L^*(x_t)$ (Berberich et al., 2021, Lemma 1), thus, resulting in practical exponential stability. The proof of this theorem as well as extensions to the case of output feedback can be found in Klöppelt et al. (2022). For a detailed discussion on the parameters λ_α, λ_σ, and c_{pe} we refer to Berberich et al. (2021).

4. CONCLUSION

In this extended abstract, we introduced a data-driven predictive control scheme, capable of stabilizing the origin even in the presence of process disturbances. To this end, we proposed a state constraint tightening which can be constructed using only system constants that can be estimated purely from data. The presented MPC scheme includes a pre-stabilizing controller and an associated input constraint tightening such that it can also cope with unstable systems. The introduced predictive controller guarantees closed-loop recursive feasibility, practical exponential stability, and constraint satisfaction.

REFERENCES

Berberich, J., Köhler, J., Müller, M.A., and Allgöwer, F. (2020a). Robust constraint satisfaction in data-driven MPC. In *Proc. 59th IEEE Conf. Decision and Control (CDC)*, 1260–1267. doi: 10.1109/CDC42340.2020.9303965.

Berberich, J., Köhler, J., Müller, M.A., and Allgöwer, F. (2021). Data-driven model predictive control with stability and robustness guarantees. *IEEE Trans. Automat. Control*, 66(4), 1702–1717. doi: 10.1109/TAC.2020.3000182.

Berberich, J., Koch, A., Scherer, C.W., and Allgöwer, F. (2020b). Robust data-driven state-feedback design. In *2020 American Control Conference (ACC)*, 1532–1538. IEEE.

Chisci, L., Rossiter, J.A., and Zappa, G. (2001). Systems with persistent disturbances: predictive control with restricted constraints. *Automatica*, 37(7), 1019–1028.

Coulson, J., Lygeros, J., and Dörfler, F. (2021). Distributionally robust chance constrained data-enabled predictive control. *IEEE Trans. Automat. Control*. Doi: 10.1109/TAC.2021.3097706.

Coulson, J., Lygeros, J., and Dörfler, F. (2019). Data-enabled predictive control: In the shallows of the DeePC. In *2019 18th European Control Conference (ECC)*, 307–312. IEEE.

Huang, L., Coulson, J., Lygeros, J., and Dörfler, F. (2021). Decentralized data-enabled predictive control for power system oscillation damping. *IEEE Trans. Control Systems Technology*. Doi: 10.1109/TCST.2021.3088638.

Klöppelt, C., Berberich, J., Allgöwer, F., and Müller, M.A. (2022). A novel constraint tightening approach for robust data-driven predictive control. *arXiv preprint arXiv:2203.07055*.

Liu, W., Sun, J., Wang, G., Bullo, F., and Chen, J. (2021). Data-driven resilient predictive control under denial-of-service. *arXiv preprint arXiv:2110.12766*.

Rawlings, J.B., Mayne, D.Q., and Diehl, M. (2017). *Model predictive control: theory, computation, and design*, volume 2. Nob Hill Publishing.

Umenberger, J. (2021). Closed-loop data-enabled predictive control. In *Proc. American Control Conf. (ACC)*, 3349–3356.

Van Waarde, H.J., Camlibel, M.K., and Mesbahi, M. (2020). From noisy data to feedback controllers: nonconservative design via a matrix S-lemma. *IEEE Trans. Automat. Control*. Doi: 10.1109/TAC.2020.3047577.

Wildhagen, S., Berberich, J., Hertneck, M., and Allgöwer, F. (2022). Data-driven analysis and controller design for discrete-time systems under aperiodic sampling. *IEEE Transactions on Automatic Control*.

Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B.L. (2005). A note on persistency of excitation. *Systems & Control Letters*, 54(4), 325–329.

Yang, H. and Li, S. (2015). A data-driven predictive controller design based on reduced hankel matrix. In *2015 10th Asian Control Conference (ASCC)*, 1–7. IEEE.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On Cameron-Liebler sets of k-spaces in finite projective spaces (Part I)

Jan De Beule *

* *Vrije Universiteit Brussel, Pleinlaan 2, B–1050 Brussel (e-mail: Jan.De.Beule@vub.be).*

Abstract: Cameron-Liebler sets of lines in a finite 3-dimensional space $\mathrm{PG}(3,q)$ originate from the study by Cameron and Liebler in 1982 of groups of collineations with equally many orbits on the points and the lines of $\mathrm{PG}(3,q)$. These objects have some interesting equivalent characterizations, and are examples of Boolean functions of degree one and completely regular codes. In this talk, we focus on these objects from a geometric perspective, and report on several existence and non-existence results, including a recent so-called modular equality for the parameter of Cameron-Liebler sets of k-spaces in finite n-dimensional projective spaces.

Keywords: low degree Boolean functions, completely regular codes, irreducible groups, Cameron-Liebler sets, extremal sets, strongly regular graphs.

1. INTRODUCTION

Let $q = p^h$ be a prime power, and denote the finite field of order q as $\mathrm{GF}(q)$. The d-dimensional projective space over $\mathrm{GF}(q)$ is the geometry consisting of all i-dimensional subspaces of the $d+1$-dimensional vector space $V(d+1,q)$ over $\mathrm{GF}(q)$, $1 \leq i \leq d$, and is denoted by $\mathrm{PG}(d,q)$. The fundamental theorem of projective geometry states that all isomorphisms of $\mathrm{PG}(d,q)$ are induced by the semi-linear maps of the underlying vector space. The group of isomorphisms of $\mathrm{PG}(d,q)$ is also called the group of collineations. Note that we call the i-dimensional subspaces of $V(d+1,q)$, $1 \leq i \leq d$, the points, lines, planes, etc. of $\mathrm{PG}(d,q)$.

In Cameron and Liebler (1982), irreducible collineation groups of $\mathrm{PG}(d,q)$, having equally many point orbits as line orbits are studied. There are several ways to characterize the line orbits of such a collineation group. One way is geometric, using spreads of the projective space. A *k-spread of* $\mathrm{PG}(d,q)$ is a set of k-dimensional projective subspaces partitioning the point set of $\mathrm{PG}(d,q)$. It is well known that a k-spread of $\mathrm{PG}(d,q)$ exists if and only if $k+1 \mid d+1$. Hence e.g. $\mathrm{PG}(d,q)$ has line spreads if and only if $2 \mid d+1$. In Cameron and Liebler (1982), it is shown that the line orbits have constant intersection with any spread, i.e. there exists a natural number x such that for any line orbit O, $|O \cap S| = x$ for any line spread S of $\mathrm{PG}(d,q)$. Such a line set O will be called a *Cameron-Liebler line class (CLLC) with parameter x.*

It was conjectured in Cameron and Liebler (1982) that such a group is line transitive or fixes a hyperplane and acts transitively on the lines of the hyperplanes, or, dually, fixes a point and acts transitively on the lines through the fixed point. Clearly, the set of lines through a fixed point p meets every spread in exactly one line. Dually, the set of lines in a fixed plane π meets every spread in exactly one line. Both line sets are CLLCs with parameter 1. They are called *trivial*. If $p \notin \pi$, the union of the set of lines through p and the set of lines in π meets every spread in exactly 2 lines. Also this example is called a trivial CLLC. Note that from the intersection property with spreads, it follows that the complement of a CLLC with parameter x in the line set, is a CLLC with parameter q^2+1-x. The complement of a trivial CLLC will also be called a trivial CLLC. The conjecture of Cameron and Liebler was disproven by Drudge (1998) and by Bruen and Drudge (1999). More precisely, in Bruen and Drudge (1999), the authors constructed a CLLC in $\mathrm{PG}(3,q)$ with parameter $x = \frac{q^2+1}{2}$ for odd q. Since the work of Bruen and Drudge, a lot of research has been done with as main objective either to construct new non-trivial CLLCs, or to show that CLLCs do not exist for particular values of the parameter x.

Now we switch to the world of low degree Boolean functions. The following theorem is well known.

Theorem 1. (Nisan and Szegedy (1994)). A Boolean degree d function on the hypercube depends on at most $d2^{d-1}$ coordinates.

In the last years, comparable results for low degree Boolean functions on different domains have been obtained. Let $J(n,k)$ denote the Johnson graph, i.e. the graph with vertex set all k-subsets of $\{1,\ldots,n\}$, two vertices being adjacent if and only if the corresponding sets meet in exactly $k-1$ elements. The following theorem is from Filmus (2016) and Meyerowitz (1992).

Theorem 2. Let $n-k, k \geq 2$. A Boolean degree 1 function on the Johnson graph $J(n,k)$ depends on at most one coordinate.

From this perspective, it is natural to consider the q-analogue of the Johnson graph as domain. This is the Grassmann graph $J_q(n,k)$, the graph with vertex set all k-dimensional subspaces of the vector space $V(n,q)$, and two vertices being adjacent if and only if their corresponding subspaces meet in a $(k-1)$-dimensional subspace of $V(n,q)$.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The original conjecture of Cameron and Liebler is translated in the language of Boolean degree one functions on $J_q(n,k)$ as follows.

Conjecture 3. (Cameron, Liebler (1982)). Let $n \geq 4$ and $k = 2$. If f is a Boolean degree 1 function on the Grassmann graph $J_q(n,k)$, then f depends on at most one point and one hyperplane.

Now we switch the viewpoint to coding theory. Let $n \geq 1$ and A be a set of q symbols. The Hamming graph $H(n,q)$ is the graph with vertex set the set of words of length n over A and two vertices being adjacent if and only if their Hamming distance is 1. The Hamming graph $H(n,2)$ is the hypercube. Clearly, a q-ary code of length n can be considered as a subset of the vertex set of $H(n,q)$. So it is quite natural to translate properties of the code C into graph theoretical properties. Conversely, it is natural to define codes as substructures in graphs different from the Hamming graph as well, replacing the Hamming distance by the graph distance. Let C be a code in a regular graph Γ with vertex set V. We follow the definition found in e.g. Neumaier (1992). Let x be any vertex of Γ, then $d(x,C) = \min\{d(x,y)|y \in C\}$. The covering radius $\rho = \max\{d(x,C)|x \in \Gamma\}$, it is the minimal integer ρ such that the spheres of radius ρ around the codewords of C cover the vertices of Γ. For a code C, minimum distance $d(C)$ and covering radius $\rho(C)$ are related by $d(C) \leq 2\rho(C)+1$. The code C is called *perfect* in case of equality, which is equivalent with the property that the spheres with radius $\rho(C)$ around the codewords partition the vertex set V.

Completely regular codes have been introduced in Delsarte (1973) as a generalization of perfect codes. Assume that C is a code in a distance regular graph. Let $C_i = \{x \in \Gamma | d(x,C) = i\}$, then $C_i \neq \emptyset \iff 0 \leq i \leq \rho(C)$, and $C_0 = C$. The sets C_i partition the vertex set of Γ. The code C is called *completely regular* if every vertex $x \in C_i$ has a constant number of neighbors a_i, b_i, respectively c_i in C_{i-1}, C_i, respectively C_{i+1}. This is equivalent with the partition of the vertex set of Γ into the components C_i being equitable.

As it is natural to move from the hypercube to the Grassmann graph as domain for low degree Boolean functions, it is equally natural to consider completely regular codes in the Grassmann graph, and it is well known that completely regular codes of covering radius 1 are equivalent to Cameron-Liebler line classes, see e.g. Filmus and Ihringer (2019).

In this talk, we will briefly overview the different points of view on CLLCs. Then we discuss the current state of the art on the existence of non-trivial CLLCs in $\mathrm{PG}(3,q)$, overview the non-existence results for CLLCs with a given particular parameter x, discuss the generalization of CLLCs to Cameron-Liebler sets of k-subspaces of the projective and affine space, and present a recent modular equality on the parameter of these generalizations.

2. KNOWN EXAMPLES AND A MODULAR EQUALITY FOR CAMERON-LIEBLER LINE CLASSES

The geometrical characterization of CLLCs, i.e. such an object has constant intersection with any spread of $\mathrm{PG}(3,q)$, is useful to generalize the notion of these objects to higher dimension if spreads exist. For d-dimensional projective spaces, this is only possible if $2 \mid d+1$. However, this restriction does not occur when dealing with affine spaces.

Recall that a d-dimensional affine space over the finite field $\mathrm{GF}(q)$ is the geometry consisting of all cosets of i-dimensional subspaces of the vector space $V(d,q)$, $0 \leq i \leq d-1$. This geometry is denoted by $\mathrm{AG}(d,q)$. It is well known that spreads of k-dimensional subspaces always exist in an affine space, it is sufficient to consider the set of k-spaces in one parallel class. A *Cameron-Liebler line set in* $\mathrm{AG}(3,q)$ is a set of lines meeting any line spread in a constant number. This is not just a straightforward generalization, as it can be shown that this definition is one of the similar characterizations possible in an affine context.

Now we are ready to overview known results for CLLCs in three dimensions.

Non-trivial examples of Cameron-Liebler line classes are rare. The first example of an infinite family was given in Bruen and Drudge (1999) (q odd, $x = \frac{q^2+1}{2}$). More recently, examples with parameter $x = \frac{q^2-1}{2}$ have been discovered in Rodgers (2013) and are described as an infinite family De Beule et al. (2016); Feng et al. (2015). These examples require $q \equiv 5$ or $9 \bmod 12$. Examples with parameter $x = \frac{(q+1)^2}{3}$ from Rodgers (2013) for $q \equiv 2 \bmod 3$ are described as an infinite family in Feng et al. (2021). Non-isomorphic derivations of some of these examples with parameter $x = \frac{q^2-1}{2}$ are found in Cossidente and Pavese (2019a) ($q > 7$ odd), and in Cossidente and Pavese (2019b) ($q \equiv 1 \bmod 4$, $q \geq 9$). Note that for some of these examples, there exists a plane of $\mathrm{PG}(3,q)$ not containing any line, and hence such examples live in $\mathrm{AG}(3,q)$. Hence, by (D'haeseleer et al., 2020, Theorem 3.8), these examples are also examples of non-trivial Cameron-Liebler line classes in $\mathrm{AG}(3,q)$. As mentioned in the introduction, non-existence results are of great interest as well, and one of the most consequential non-existence conditions is the following theorem.

Theorem 4. (Gavrilyuk and Metsch, 2014, Theorem 1.1) Suppose that L is a Cameron-Liebler line class with parameter x of $\mathrm{PG}(3,q)$. Then for every plane and every point of $\mathrm{PG}(3,q)$,

$$\binom{x}{2} + m(m-x) \equiv 0 \bmod (q+1), \tag{1}$$

where m is the number of lines of L in the plane, respectively through the point.

As a corollary of Theorem 4, the following theorem is given in D'haeseleer et al. (2020)

Theorem 5. Suppose that L is a Cameron-Liebler line class in $\mathrm{AG}(3,q)$ with parameter x. Then

$$x(x-1) \equiv 0 \bmod 2(q+1)\,. \tag{2}$$

3. A MODULAR EQUALITY FOR CAMERON-LIEBLER SETS OF K-SPACES

The main results we present in the talk, is the generalization of Theorem 4, respectively 5, to Cameron-Liebler

line classes in projective spaces, respectively affine spaces of dimension $n \geq 3$ odd. This is joint work with Jonathan Mannaert.

Recall the Gaussian binomial coefficient, for $a \leq b$ natural numbers,

$$\begin{bmatrix} b \\ a \end{bmatrix}_q = \frac{(q^b-1)\dots(q^{b-a+1}-1)}{(q^a-1)\dots(q-1)}$$

which represents the number of $(a-1)$-dimensional projective spaces in a projective space of dimension $b-1$.

For the sake of completeness, we mention one alternative definition of Cameron-Liebler sets of lines in $\mathrm{PG}(d,q)$ that avoids the use of line spreads, which do not exist if d is even. A *Cameron-Liebler set of lines in* $\mathrm{PG}(d,q)$ is a set L of lines of which the characteristic vector χ_L is a linear combination of characteristic vectors of the point-line pencils in $\mathrm{PG}(d,q)$. In case $d=3$, then the number

$$x = \frac{|L|}{\begin{bmatrix} d \\ 1 \end{bmatrix}_q}$$

coincides with the parameter of L as defined in Section 1. Therefore we use this number as definition of the *parameter of a Cameron-Liebler set of lines in* $\mathrm{PG}(d,q)$ for $d \geq 3$. The parameter will be an integer if and only if d is odd.

The following lemma originates from Drudge (1998). It turns out to be essential to prove the desired generalization of Theorem 4. A short proof can e.g. be found in De Beule et al. (2022).

Lemma 6. Let L be a Cameron-Liebler set of lines in $\mathrm{PG}(d,q)$, $d \geq 3$, and let π be any i-dimensional projective subspace of $\mathrm{PG}(d,q)$, $i \geq 2$. Then the set of lines of L contained in π is a Cameron-Liebler set of lines in π, with parameter x_π.

Secondly, a series of combinatorial lemmas are needed. These are the following.

Lemma 7. (De Beule et al., 2022, Theorem 5.1 for $k=1$ and $t=3$) Suppose that $\mathcal{L}$ is a non-empty Cameron-Liebler line class in $\mathrm{PG}(n,q)$, $n \geq 4$ even, with parameter x. Then

$$x = 1 + \frac{C}{\begin{bmatrix} n-2 \\ 1 \end{bmatrix}_q},$$

for some $C \in \mathbb{N}$.

Lemma 8. (Blokhuis et al., 2019, Theorem 2.9) Suppose that $\mathcal{L}$ is a Cameron- Liebler line class with parameter x in $\mathrm{PG}(n,q)$, with $n \geq 3$. If ℓ is an arbitrary line in $\mathrm{PG}(n,q)$ then there are in total $q^2\frac{q^{n-2}-1}{q-1}(x-\chi(\ell))$ lines of $\mathcal{L}$ skew to ℓ. Here $\chi(\ell)$ equals one if $\ell \in \mathcal{L}$ or zero otherwise.

Lemma 9. (Segre, 1961, Section 170) The number of j-spaces disjoint to a fixed m-space in $\mathrm{PG}(n,q)$ is equal to $q^{(m+1)(j+1)} \begin{bmatrix} n-m \\ j+1 \end{bmatrix}_q$.

The arguments to prove the following theorem, are of combinatorial nature. The above Lemmas 6, 7, 8, and 9 enable a series of arguments that relate the parameter of a Cameron-Liebler set of lines to the parameter of the induced Cameron-Liebler line class in different 3-dimensional projective spaces, the latter for which we can use Theorem 4. The main result for the projective case is the following theorem.

Theorem 10. (De Beule and Mannaert (2022)). Let L be a Cameron-Liebler line class with parameter x in $\mathrm{PG}(n,q)$, with $n \geq 7$ odd. Then for any point p,

$$x(x-1) + 2\overline{m}(\overline{m}-x) \equiv 0 \mod (q+1),$$

where $\overline{m}$ is the number of lines of L through p.

Lemma 6 can be reformulated for Cameron-Liebler line sets in affine spaces, see e.g. De Beule et al. (2022).

Lemma 11. Suppose that $\mathcal{L}$ is a Cameron-Liebler set of lines in $\mathrm{AG}(n,q)$, $n >$. Then for every i-dimensional subspace π, with $i > 1$ the set of lines of L in π is a Cameron-Liebler line set in π, with parameter x_π.

The generalization of Theorem 5 now becomes

Theorem 12. (De Beule and Mannaert (2022)). Let L be a Cameron-Liebler line class in $\mathrm{AG}(n,q)$, $n \geq 3$ odd, with parameter x, then

$$x(x-1) \equiv 0 \mod 2(q+1).$$

We will illustrate how both Theorem 10 and 12 can be used to reduce the admitted parameters for given n and q.

REFERENCES

Blokhuis, A., De Boeck, M., and D'haeseleer, J. (2019). Cameron-Liebler sets of k-spaces in $\mathrm{PG}(n,q)$. *Des. Codes Cryptogr.*, 87(8), 1839–1856. doi:10.1007/s10623-018-0583-1. URL https://doi.org/10.1007/s10623-018-0583-1.

Bruen, A.A. and Drudge, K. (1999). The construction of Cameron-Liebler line classes in $\mathrm{PG}(3,q)$. *Finite Fields Appl.*, 5(1), 35–45. doi:10.1006/ffta.1998.0239. URL https://doi.org/10.1006/ffta.1998.0239.

Cameron, P.J. and Liebler, R.A. (1982). Tactical decompositions and orbits of projective groups. *Linear Algebra Appl.*, 46, 91–102. doi:10.1016/0024-3795(82)90029-5. URL https://doi.org/10.1016/0024-3795(82)90029-5.

Cossidente, A. and Pavese, F. (2019a). New Cameron-Liebler line classes with parameter $\frac{q^2+1}{2}$. *J. Algebraic Combin.*, 49(2), 193–208. doi:10.1007/s10801-018-0826-2. URL https://doi.org/10.1007/s10801-018-0826-2.

Cossidente, A. and Pavese, F. (2019b). Cameron-Liebler line classes of $\mathrm{PG}(3,q)$ admitting $\mathrm{PGL}(2,q)$. *J. Combin. Theory Ser. A*, 167, 104–120. doi:10.1016/j.jcta.2019.04.004. URL https://doi.org/10.1016/j.jcta.2019.04.004.

De Beule, J. and Mannaert, J. (2022). A modular equality for Cameron-Liebler line classes in projective and affine spaces of odd dimension. *Finite Fields Appl.*, 82, Paper No. 102047. doi:10.1016/j.ffa.2022.102047. URL https://doi.org/10.1016/j.ffa.2022.102047.

De Beule, J., Demeyer, J., Metsch, K., and Rodgers, M. (2016). A new family of tight sets in $Q^+(5,q)$. *Des. Codes Cryptogr.*, 78(3), 655–678. doi:10.1007/s10623-014-0023-9. URL https://doi.org/10.1007/s10623-014-0023-9.

De Beule, J., Mannaert, J., and Storme, L. (2022). Cameron-Liebler k-sets in subspaces and non-existence conditions. *Des. Codes Cryptogr.*, 90(3),

633–651. doi:10.1007/s10623-021-00995-0. URL https://doi.org/10.1007/s10623-021-00995-0.

Delsarte, P. (1973). An algebraic approach to the association schemes of coding theory. *Philips Res. Rep. Suppl.*, (10), vi+97.

D'haeseleer, J., Mannaert, J., Storme, L., and Švob, A. (2020). Cameron-Liebler line classes in AG$(3, q)$. *Finite Fields Appl.*, 67, 101706, 17. doi:10.1016/j.ffa.2020.101706. URL https://doi.org/10.1016/j.ffa.2020.101706.

Drudge, K.W. (1998). *Extremal sets in projective and polar spaces*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–The University of Western Ontario (Canada).

Feng, T., Momihara, K., Rodgers, M., Xiang, Q., and Zou, H. (2021). Cameron-Liebler line classes with parameter $x = \frac{(q+1)^2}{3}$. *Adv. Math.*, 385, Paper No. 107780, 31. doi:10.1016/j.aim.2021.107780. URL https://doi.org/10.1016/j.aim.2021.107780.

Feng, T., Momihara, K., and Xiang, Q. (2015). Cameron-Liebler line classes with parameter $x = \frac{q^2-1}{2}$. *J. Combin. Theory Ser. A*, 133, 307–338. doi:10.1016/j.jcta.2015.02.004. URL https://doi.org/10.1016/j.jcta.2015.02.004.

Filmus, Y. (2016). Friedgut-Kalai-Naor theorem for slices of the Boolean cube. *Chic. J. Theoret. Comput. Sci.*, Art. 14, 17. doi:10.4086/cjtcs.2016.014. URL https://doi.org/10.4086/cjtcs.2016.014.

Filmus, Y. and Ihringer, F. (2019). Boolean degree 1 functions on some classical association schemes. *J. Combin. Theory Ser. A*, 162, 241–270. doi:10.1016/j.jcta.2018.11.006. URL https://doi.org/10.1016/j.jcta.2018.11.006.

Gavrilyuk, A.L. and Metsch, K. (2014). A modular equality for Cameron-Liebler line classes. *J. Combin. Theory Ser. A*, 127, 224–242. doi:10.1016/j.jcta.2014.06.004. URL https://doi.org/10.1016/j.jcta.2014.06.004.

Meyerowitz, A.D. (1992). Cycle-balanced partitions in distance-regular graphs. *J. Combin. Inform. System Sci.*, 17(1-2), 39–42.

Neumaier, A. (1992). Completely regular codes. volume 106/107, 353–360. doi:10.1016/0012-365X(92)90565-W. URL https://doi.org/10.1016/0012-365X(92)90565-W. A collection of contributions in honour of Jack van Lint.

Nisan, N. and Szegedy, M. (1994). On the degree of Boolean functions as real polynomials. volume 4, 301–313. doi:10.1007/BF01263419. URL https://doi.org/10.1007/BF01263419. Special issue on circuit complexity (Barbados, 1992).

Rodgers, M. (2013). Cameron-Liebler line classes. *Des. Codes Cryptogr.*, 68(1-3), 33–37. doi:10.1007/s10623-011-9581-2. URL https://doi.org/10.1007/s10623-011-9581-2.

Segre, B. (1961). *Lectures on modern geometry*, volume 7 of *Consiglio Nazionale delle Ricerche Monografie Matematiche*. Edizioni Cremonese, Rome. With an appendix by Lucio Lombardo-Radice.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Global boundary stabilization of a semilinear heat equation via finite-dimensional nonlinear observers

Rami Katz Emilia Fridman

* *School of Electrical Engineering, Tel-Aviv University, Tel-Aviv*
(e-mail: ramikatz@mail.tau.ac.il,emilia@tauex.tau.ac.il)

Abstract: We study global finite-dimensional observer-based stabilization of a 1D heat equation with a known globally Lipschitz semilinearity in the state variable. We consider Neumann actuation and point measurement. Using dynamic extension and modal decomposition we derive nonlinear ODEs for the modes of the state. We then design a finite-dimensional nonlinear Luenberger observer, which takes into account the known semilinearity. The proposed controller is based on this observer. Our Lypunov H^1-stability analysis leads to LMIs, which are feasible for a large enough observer dimension and small enough Lipschitz constant.

Keywords: Distributed parameter systems, nonlinear systems, observer-based control, Lyapunov method.

1. INTRODUCTION

Observer-based control of parabolic PDEs is a challenging problem with numerous applications, including chemical reactors, flame propagation and viscous flow (Christofides [2001]). Output-feedback controllers for PDEs have been constructed by the modal decomposition approach (Curtain [1982], Lasiecka and Triggiani [2000], Orlov et al. [2004]), the backstepping method (Krstic and Smyshlyaev [2008]) and the spatial decomposition approach (Fridman and Blighovsky [2012], Kang and Fridman [2020]). Constructive finite-dimensional observer-based design for linear 1D parabolic PDEs was introduced in (Katz and Fridman [2020, 2021a]), via modal decomposition. The challenging problem of efficient finite-dimensional observer-based design for semilinear parabolic PDEs remained open.

State-feedback control of some semilinear PDEs was studied in (Vazquez and Krstic [2008]) using backstepping, in (Karafyllis and Krstic [2019]) using small-gain theorem and in (Karafyllis [2021]) via control Lyapunov functions. Recently, modal-decomposition-based state-feedback was proposed in (Katz and Fridman [2021b]) for global stabilization of heat equation and in (Katz and Fridman [2021a]) for regional stabilization of Kuramoto-Sivashinsky equation. Finite-dimensional control based on linear observers was proposed in (Wu et al. [2016]) for semilinear parabolic PDEs via modal decomposition. Linear observers should have high gains required to dominate the nonlinearity, which leads to small delays that preserve the stability (Lei and Khalil [2016], Najafi and Ekramian [2021]).

For semilinear parabolic PDEs, efficient finite-dimensional observer-based controller design remained an open challenging problems that we address in the present paper. We consider global stabilization of a semilinear heat equation under Neumann actuation and point measurement. The semilinarity is assumed to be globally Lipschitz in the state. Using dynamic extension and modal decomposition we derive nonlinear ODEs for the modes of the state. We then design a finite-dimensional *nonlinear* Luenberger observer, which takes into account the known semilinearity. The proposed controller is based on the nonlinear finite-dimensional observer. The challenge in the Lyapunov-based analysis is due to the coupling between the finite-dimensional and infinite-dimensional parts of the closed-loop system, introduced by both the semilinearity and the estimation error. Our H^1-stability analysis leads to LMIs, which are feasible for a large enough observer dimension and small enough Lipschitz constant. The results in this manuscript have been recently extended to finite-dimensional observer-based stabilization of a semilinear heat in the presence of large input delay in (Katz and Fridman [2022]).

Preliminaries: $L^2(0,1)$ is the space of Lebesgue measurable and square integrable functions $f:[0,1]\to\mathbb{R}$ with the inner product $\langle f,g\rangle := \int_0^1 f(x)g(x)dx$ and induced norm $\|f\|^2 := \langle f,f\rangle$. $H^k(0,1)$ is the Sobolev space of functions $f:[0,1]\to\mathbb{R}$ having k square integrable weak derivatives, with the norm $\|f\|_{H^k}^2 := \sum_{j=0}^k \left\|f^{(j)}\right\|^2$. The Euclidean norm on $\mathbb{R}^n$ is denoted by $|\cdot|$. For $P\in\mathbb{R}^{n\times n}$, $P>0$ means that P is symmetric and positive definite. The subdiagonal elements of a symmetric matrix will be denoted by $*$.

Consider the Sturm-Liouville eigenvalue problem

$$\phi''+\lambda\phi=0,\quad x\in(0,1) \tag{1.1}$$

with boundary conditions

$$\phi'(0)=\phi'(1)=0. \tag{1.2}$$

⋆ Supported by Israel Science Foundation (grant no. 673/19) and by C. and H. Manderman Chair at Tel Aviv University

This problem induces a sequence of eigenvalues with corresponding eigenfunctions. The normalized eigenfunctions form a complete orthonormal system in $L^2(0,1)$. The eigenvalues and corresponding eigenfunctions are given by

$$\begin{aligned}&\phi_0(x)\equiv 1,\ \phi_n(x)=\sqrt{2}\cos\left(\sqrt{\lambda_n}x\right),\\&\lambda_n=n^2\pi^2,\ n\in\mathbb{Z}_+.\end{aligned}\tag{1.3}$$

The following lemmas will be used:

Lemma 1.1. (Katz and Fridman [2020]) Let $h\overset{L^2}{=}\sum_{n=0}^{\infty}h_n\phi_n$. Then $h\in H^2(0,1)$ with $h'(0)=h'(1)=0$ if and only if $\sum_{n=1}^{\infty}\lambda_n^2h_n^2<\infty$. Moreover, $\|h'\|^2=\sum_{n=1}^{\infty}\lambda_nh_n^2$.

Lemma 1.2. (Sobolev's inequality, Kang and Fridman [2019]) Let $h\in H^1(0,1)$. Then, for all $\Gamma>0$, $max_{x\in[0,1]}|h(x)|^2\le(1+\Gamma)\|h\|^2+\Gamma^{-1}\|h'\|^2$.

2. FINITE-DIMENSIONAL OBSERVER-BASED CONTROL OF A SEMILINEAR HEAT EQUATION

We consider stabilization of the semilinear 1D heat equation

$$z_t(x,t)=z_{xx}(x,t)+g\left(t,x,z(x,t)\right),\ t\ge 0\tag{2.1}$$

where $x\in[0,1]$ and $z(x,t)\in\mathbb{R}$. We consider Neumann actuation

$$z_x(0,t)=0,\quad z_x(1,t)=u(t)\tag{2.2}$$

where $u(t)$ is a control input to be designed. We further assume point measurement given by

$$y(t)=z(x_*,t),\quad x_*\in[0,1].\tag{2.3}$$

Note that $x_*=0$ or $x_*=1$ correspond to boundary measurements. Here $g:\mathbb{R}^3\to\mathbb{R}$ is a locally Lipschitz function which satisfies $g(t,x,0)\equiv 0$ and

$$\sup_{z_1\ne z_2}\frac{|g(t,x,z_1)-g(t,x,z_2)|}{|z_1-z_2|}\le\sigma,\ \forall\,(t,x)\in\mathbb{R}^2\tag{2.4}$$

for some $\sigma>0$, independent of $(t,x)\in\mathbb{R}^2$.

Let $\psi(x)=-\frac{2}{\pi}\cos\left(\frac{\pi}{2}x\right)$ and note that it satisfies

$$\begin{aligned}&\psi''(x)=-\mu\psi(x),\quad \mu=\frac{\pi^2}{4},\\&\psi'(0)=0,\ \psi'(1)=1,\ \|\psi\|^2=\frac{2}{\pi^2}.\end{aligned}\tag{2.5}$$

Forthermore,

$$\langle\psi,\phi_0\rangle=\frac{4}{\pi^2},\ \langle\psi,\phi_n\rangle=\frac{\sqrt{2}(-1)^n}{\lambda_n-\mu},\ n\ge 1.\tag{2.6}$$

Similar to (Karafyllis [2021]), we introduce

$$w(x,t)=z(x,t)-\psi(x)u(t).\tag{2.7}$$

We define further the new control input $v(t)$ that satisfies the following relations:

$$\dot u(t)=-\mu u(t)+v(t),\quad u(0)=0,\quad t\ge 0.$$

Then we obtain the equivalent ODE-PDE system

$$\begin{aligned}&\dot u(t)=-\mu u(t)+v(t),\quad t\ge 0,\\&w_t(x,t)=w_{xx}(x,t)+g\left(t,x,w(x,t)+\psi(x)u(t)\right)\\&\qquad\quad -\psi(x)v(t),\\&w_x(0,t)=w_x(1,t)=0\end{aligned}\tag{2.8}$$

with measurement

$$y(t)=w(x_*,t)+\psi(x_*)u(t).\tag{2.9}$$

We will treat further $u(t)$ as an additional state variable.

We present the solution to (2.8) as

$$w(x,t)=\sum_{n=0}^{\infty}w_n(t)\phi_n(x),\ w_n(t)=\langle w(\cdot,t),\phi_n\rangle,\tag{2.10}$$

with $\{\phi_n\}_{n=0}^{\infty}$ defined in (1.3). By differentiating under the integral sign, integrating by parts and using (1.1) and (1.2) we obtain for $t\ge 0$

$$\begin{aligned}&\dot w_n(t)=-\lambda_nw_n(t)+g_n(t)+b_nv(t),\\&w_n(0)=\langle w(\cdot,0),\phi_n\rangle,\end{aligned}\tag{2.11}$$

where

$$\begin{aligned}&g_n(t)=\langle g\left(t,\cdot,w(\cdot,t)+\psi(\cdot)u(t)\right),\phi_n\rangle,\\&b_0\overset{(2.6)}{=}\frac{4}{\pi^2},\quad b_n\overset{(2.6)}{=}\frac{(-1)^{n+1}4\sqrt{2}}{\pi^2\left(4n^2-1\right)},\ n\ge 1.\end{aligned}\tag{2.12}$$

Note that given $N\in\mathbb{Z}_+$, (2.12) and the integral test for series convergence imply

$$\begin{aligned}&\sum_{n=N+1}^{\infty}\lambda_nb_n^2=\frac{32}{\pi^2}\sum_{n=N+1}^{\infty}\frac{n^2}{(4n^2-1)^2}\le\frac{2\xi_{N+1}}{\pi^2},\\&\xi_{N+1}=\left(1+\frac{1}{4(N+1)^2-1}\right)^2\frac{1}{N}.\end{aligned}\tag{2.13}$$

Let $\delta>0$ be a desired decay rate and let $N_0\in\mathbb{Z}_+$ satisfy

$$-\lambda_n+\sigma<-\delta,\quad n>N_0.\tag{2.14}$$

N_0 is the number of modes in our controller, whereas $N\in\mathbb{Z}_+$, $N\ge N_0$ is the observer dimension. We construct a finite-dimensional observer of the form

$$\hat w(x,t)=\sum_{n=0}^{N}\hat w_n(t)\phi_n(x)\tag{2.15}$$

where $\{\hat w_n(t)\}_{n=0}^{N}$ satisfy the *nonlinear* ODEs

$$\begin{aligned}&\dot{\hat w}_n(t)=-\lambda_n\hat w_n(t)+\hat g_n(t)+b_nv(t)\\&-l_n\left[\hat w(x_*,t)+\psi(x_*)u(t)-y(t)\right],\ 0\le n\le N\end{aligned}\tag{2.16}$$

with scalar observer gains $\{l_n\}_{n=0}^{N}$ and

$$\hat g_n(t)=\langle g\left(t,\cdot,\hat w(\cdot,t)+\psi(\cdot)u(t)\right),\phi_n\rangle,\ 0\le n\le N.\tag{2.17}$$

In particular, we approximate the projections of the semilinearity $g(t,x,w(x,t)+\psi(x)u(t))$ onto $\{\phi_n\}_{n=0}^{N}$ by the projections of $g(t,x,\hat w(x,t)+\psi(x)u(t))$ onto $\{\phi_n\}_{n=0}^{N}$.

Assumption 1: The point $x_*\in[0,1]$ satisfies

$$c_n=\phi_n(x_*)\ne 0,\quad 0\le n\le N_0.\tag{2.18}$$

Note that Assumption 1 holds for the particular case of boundary measurements $x_*=0$ or $x_*=1$.

Denote

$$\begin{aligned}&\tilde A_0=\mathrm{diag}\left\{-\mu,A_0\right\},\quad \tilde B_0=\mathrm{col}\left\{1,B_0\right\}\\&A_0=\mathrm{diag}\left\{-\lambda_n\right\}_{n=0}^{N_0},\quad B_0=\mathrm{col}\left\{b_n\right\}_{n=0}^{N_0}\\&C_0=[c_0,\ldots,c_{N_0}],\quad C_1=[c_{N_0+1},\ldots,c_N],\end{aligned}\tag{2.19}$$

Under Assumption 1, the pair (A_0,C_0) is observable by the Hautus lemma. Let $L_0=\{l_n\}_{n=0}^{N_0}\in\mathbb{R}^{N_0+1}$ satisfy the Lyapunov inequality

$$P_{\mathrm{o}}(A_0-L_0C_0)+(A_0-L_0C_0)^TP_{\mathrm{o}}<-2\delta P_{\mathrm{o}}\tag{2.20}$$

with $0<P_{\mathrm{o}}\in\mathbb{R}^{(N_0+1)\times(N_0+1)}$. We further choose the remaining gains as $l_n=0,\ N_0+1\le n\le N$.

Similarly, by the Hautus lemma, the pair $(\tilde A_0,\tilde B_0)$ is controllable. Let $K_0\in\mathbb{R}^{1\times(N_0+2)}$ satisfy

$$P_{\mathrm{c}}(\tilde A_0-\tilde B_0K_0)+(\tilde A_0-\tilde B_0K_0)^TP_{\mathrm{c}}<-2\delta P_{\mathrm{c}},\tag{2.21}$$

with $0 < P_{\mathrm{c}} \in \mathbb{R}^{(N_0+2)\times(N_0+2)}$. We propose the controller

$$v(t) = -K_0\hat{w}^{N_0}(t),\ \hat{w}^{N_0}(t) = \mathrm{col}\left\{u(t), \hat{w}_n(t)\right\}_{n=0}^{N_0} \tag{2.22}$$

which is based on the finite-dimensional observer (2.15).

For well-posedness of the closed-loop system (2.7), (2.16) subject to the control law (2.22), consider the operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to L^2(0,1)$, $\mathcal{A} = -\partial_{xx}$ with $\mathcal{D}(\mathcal{A}) = \left\{h \in H^2(0,1) \mid h'(0) = h'(1) = 0\right\}$. Let $\theta > 0$ and $\mathcal{A}_\theta = \mathcal{A} + \theta I$. $-\mathcal{A}_\theta$ generates an analytic semigroup on $L^2(0,1)$. Moreover, there exists a unique positive root $\mathcal{A}_\theta^{\frac{1}{2}}$ with $\mathcal{D}\left(\mathcal{A}_\theta^{\frac{1}{2}}\right) = H^1(0,1)$. Let $\mathcal{H} = L^2(0,1)\times\mathbb{R}^{N+2}$ be a Hilbert space with the norm $\|\cdot\|_{\mathcal{H}}^2 := \|\cdot\|^2 + |\cdot|^2$. Introducing

$$\begin{aligned}&\xi(t) = \mathrm{col}\left\{\xi_1(t), \xi_2(t)\right\},\ \xi_1(t) = w(\cdot,t),\ \xi_2(t) = \hat{w}^N(t),\\ &\hat{w}^N(t) = \mathrm{col}\left\{u(t), \hat{w}_0(t), \dots, \hat{w}_N(t)\right\}\end{aligned} \tag{2.23}$$

the closed-loop system can be presented as

$$\begin{aligned}&\frac{d\xi}{dt}(t) + \mathrm{diag}\left\{\mathcal{A}_\theta, \mathcal{B}\right\}\xi(t) = \begin{bmatrix} f_1(\xi) \\ f_2(\xi)\end{bmatrix},\\ &\mathcal{D}(\mathcal{B}) = \mathbb{R}^{N+2},\ \mathcal{B}a = \begin{bmatrix} -\tilde{A}_0 + \tilde{B}_0K_0 + \tilde{L}_0[0\ C_0] & \tilde{L}_0C_1 \\ B_1K_0 & -A_1\end{bmatrix} a\\ &f_1(t,\xi) = \theta w(\cdot,t) + g\left(t,\cdot,w(\cdot,t) + \psi(\cdot)u(t)\right)\\ &\qquad +\psi(\cdot)K_0\hat{w}^{N_0}(t),\\ &f_2(t,\xi) = \mathrm{col}\left\{\hat{G}^{N_0}(t) + \tilde{L}_0 w(x_*,t), \hat{G}^{N-N_0}(t)\right\},\\ &\hat{G}^{N_0}(t) = \mathrm{col}\left\{0, \hat{g}_n(t)\right\}_{n=0}^{N_0},\\ &\hat{G}^{N-N_0}(t) = \mathrm{col}\left\{\hat{g}_n(t)\right\}_{n=N_0+1}^{N},\ \tilde{L}_0 = \mathrm{col}\left\{0, l_n\right\}_{n=0}^{N_0},\\ &A_1 = \mathrm{diag}\left\{-\lambda_n\right\}_{n=N_0+1}^{N},\ B_1 = \mathrm{col}\left\{b_n\right\}_{n=N_0+1}^{N}.\end{aligned} \tag{2.24}$$

It can be shown that given $w(\cdot,0) \in H^1(0,1)$, the system (2.24) has a unique classical solution satisfying $\xi \in C\left([0,\infty);\mathcal{H}\right) \cap C^1\left((0,\infty);\mathcal{H}\right)$ and such that $\xi(t) \in \mathcal{D}\left(\mathrm{diag}\left\{\mathcal{A}_\theta, \mathcal{B}\right\}\right) = \mathcal{D}(\mathcal{A}) \times \mathbb{R}^{N+2}\ \ \forall t > 0$. Details are omitted due to space constraints (see (Katz and Fridman [2022])).

Introduce the estimation error $e_n(t) = w_n(t) - \hat{w}_n(t)$, $0 \le n \le N_0$. Using the estimation error and $\{c_n\}_{n=0}^N$ in (2.19), the innovation term in (2.16) can be presented as

$$\begin{aligned}&\hat{w}(x_*,t) + \psi(x_*)u(t) - y(t) = \hat{w}(x_*,t) - w(x_*,t)\\ &= -\sum_{n=0}^{N} c_n e_n(t) - \zeta(t),\ \zeta(t) = w(0,t) - \sum_{n=0}^{N} w_n(t)\phi_n.\end{aligned} \tag{2.25}$$

Let $\Gamma > 0$. By Lemmas 1.1 and 1.2 we have

$$\zeta^2(t) \le \sum_{n=N+1}^{\infty} \kappa_n w_n^2(t),\ \kappa_n = 1 + \Gamma + \Gamma^{-1}\lambda_n. \tag{2.26}$$

Taking into account (2.11), (2.16), (2.19) and (2.25), the estimation error satisfies the following ODEs

$$\begin{aligned}&\dot{e}_n(t) = -\lambda_n e_n(t) + h_n(t)\\ &\qquad -l_n\sum_{n=0}^{N} c_n e_n(t) - l_n\zeta(t),\ 0 \le n \le N_0,\\ &\dot{e}_n(t) = -\lambda_n e_n(t) + h_n(t),\ N_0 + 1 \le n \le N.\end{aligned} \tag{2.27}$$

where we define

$$h_n(t) = g_n(t) - \hat{g}_n(t), \quad n \ge 0. \tag{2.28}$$

Recall (2.19) and denote

$$\begin{aligned}&\hat{w}^{N-N_0}(t) = \mathrm{col}\left\{\hat{w}_n(t)\right\}_{n=N_0+1}^{N},\ e^{N_0}(t) = \mathrm{col}\left\{e_n(t)\right\}_{n=0}^{N_0},\\ &e^{N-N_0}(t) = \mathrm{col}\left\{e_n(t)\right\}_{n=N_0+1}^{N},\ H^{N_0}(t) = \mathrm{col}\left\{h_n(t)\right\}_{n=0}^{N_0},\\ &H^{N-N_0}(t) = \mathrm{col}\left\{h_n(t)\right\}_{n=N_0+1}^{N},\ L_\zeta = \mathrm{col}\left\{\tilde{L}_0, -L_0, 0, 0\right\},\\ &X(t) = \mathrm{col}\left\{\hat{w}^{N_0}(t), e^{N_0}(t), \hat{w}^{N-N_0}(t), e^{N-N_0}(t)\right\},\\ &\hat{G}(t) = \mathrm{col}\left\{\hat{G}^{N_0}(t), 0, \hat{G}^{N-N_0}(t), 0\right\},\ K_X = [K_0, 0, 0, 0],\\ &H(t) = \mathrm{col}\left\{0, H^{N_0}(t), 0, H^{N-N_0}(t)\right\}.\end{aligned} \tag{2.29}$$

Using (2.11), (2.16) - (2.19), (2.22), (2.25), (2.27) and (2.29), the closed-loop system for $t \ge 0$ is presented as

$$\begin{aligned}&\dot{X}(t) = F_X X(t) + L_\zeta\zeta(t) + \hat{G}(t) + H(t),\\ &\dot{w}_n(t) = -\lambda_n w_n(t) + \hat{g}_n(t) + h_n(t)\\ &\qquad -b_n K_X X(t),\ n > N,\\ &F_X = \begin{bmatrix} \tilde{A}_0 - \tilde{B}_0K_0 & \tilde{L}_0C_0 & 0 & \tilde{L}_0C_1 \\ 0 & A_0 - L_0C_0 & 0 & -L_0C_1 \\ -B_1K_0 & 0 & A_1 & 0 \\ 0 & 0 & 0 & A_1\end{bmatrix}.\end{aligned} \tag{2.30}$$

For H^1-stability analysis of the closed-loop system (2.30) we consider the Lyapunov function

$$V(t) = X^T(t)P_X X(t) + \sum_{n=N+1}^{\infty} \lambda_n w_n^2(t) \tag{2.31}$$

where $0 < P_X \in \mathbb{R}^{(2N+3)\times(2N+3)}$. Differentiating $V(t)$ along the solution to the closed-loop system (2.30) we have

$$\begin{aligned}&\dot{V} + 2\delta V = 2X^T(t)\left[P_X F_X + F_X^T P_X + 2\delta P_X\right]X(t)\\ &+2X^T(t)P_X L_\zeta\zeta(t) + 2X^T(t)P_X\hat{G}(t) + 2X^T(t)P_X H(t)\\ &+2\sum_{n=N+1}^{\infty}\left(-\lambda_n^2 + \delta\lambda_n\right)w_n^2(t)\\ &+2\sum_{n=N+1}^{\infty}\lambda_n w_n(t)\left[\hat{g}_n(t) + h_n(t) - b_N K_X X(t)\right].\end{aligned} \tag{2.32}$$

Let $\alpha_1 > 0$, we compensate the series with $\{\hat{g}_n(t)\}_{n=N+1}^{\infty}$ by using the Young inequality

$$\begin{aligned}&2\sum_{n=N+1}^{\infty}\lambda_n w_n(t)\hat{g}_n(t) \le \frac{1}{\alpha_1}\sum_{n=N+1}^{\infty}\lambda_n^2 w_n^2(t)\\ &-\alpha_1\left|\hat{G}(t)\right|^2 + \alpha_1\sum_{n=0}^{\infty}\hat{g}_n^2(t),\\ &\alpha_1\sum_{n=0}^{\infty}\hat{g}_n^2(t) \overset{(2.4)}{\le} 2\alpha_1\sigma^2 X^T(t)\Xi_X X(t)\\ &\Xi_X \overset{(2.5)}{=} \mathrm{diag}\left\{\frac{2}{\pi^2}, I_{N_0+1}, 0, I_{N-N_0}, 0\right\}.\end{aligned} \tag{2.33}$$

Similarly, introducing $\alpha_2 > 0$ we have

$$\begin{aligned}&2\sum_{n=N+1}^{\infty}\lambda_n w_n(t)h_n(t) \le \frac{1}{\alpha_2}\sum_{n=N+1}^{\infty}\lambda_n^2 w_n^2(t)\\ &-\alpha_2\left|H(t)\right|^2 + \alpha_2\sigma^2 X^T(t)\Xi_E X(t) + \alpha_2\sigma^2\sum_{n=N+1}^{\infty} w_n^2(t),\\ &\Xi_E = \mathrm{diag}\left\{0, I_{N_0}, 0, I_{N-N_0}\right\} \in \mathbb{R}^{(2N+3)\times(2N+3)}.\end{aligned} \tag{2.34}$$

We bound the last term in (2.32) by using Young's inequality with some $\alpha_3 > 0$:

$$2\sum_{n=N+1}^{\infty}\lambda_n w_n(t)\left(-b_n K_X X(t)\right)\overset{(2.13)}{\leq}\frac{1}{\alpha_3}\sum_{n=N+1}^{\infty}\lambda_n w_n^2(t)+\frac{2\alpha_3\xi_{N+1}}{\pi^2}\left|K_X X(t)\right|^2. \tag{2.35}$$

Let $\rho_n=\kappa_n^{-1}\left(-\lambda_n^2+\delta\lambda_n+\frac{\lambda_n}{2\alpha_3}+\frac{\lambda_n^2}{2\alpha_2}+\frac{\lambda_n^2}{2\alpha_2}+\frac{\alpha_2\sigma^2}{2}\right)$ for $n\geq N$. Assuming that $\rho_{N+1}<0$, it can be seen that ρ_n is monotonically decreasing. The latter follows from monotonicity of λ_n. Then

$$\sum_{n=N+1}^{\infty}\left(-\lambda_n^2+\delta\lambda_n+\frac{\lambda_n}{2\alpha_3}+\frac{\lambda_n^2}{2\alpha_1}+\frac{\lambda_n^2}{2\alpha_2}+\frac{\alpha_2\sigma^2}{2}\right)w_n^2(t)=\sum_{n=N+1}^{\infty}\rho_n\kappa_n w_n^2(t)\overset{(2.26)}{\leq}\rho_{N+1}\zeta^2(t). \tag{2.36}$$

Let $\eta(t)=\operatorname{col}\left\{X(t),\zeta(t),\hat{G}(t),H(t)\right\}$. From (2.32)-(2.36) we have $\dot{V}+2\delta V\leq\eta^T(t)\Psi_0\eta(t)\leq 0$, provided

$$\Psi_0=\left[\begin{array}{cc|cc}\psi_0 & P_X L_\zeta & P_X & P_X\\ * & 2\rho_{N+1} & 0 & 0\\ \hline * & & \multicolumn{2}{c}{\operatorname{diag}\{-\alpha_1 I,-\alpha_2 I\}}\end{array}\right]<0,\ \psi_0=P_X F_X$$
$$F_X^T P_X+2\delta P_X+\frac{2\alpha_3\xi_{N+1}}{\pi^2}K_X^T K_X+2\alpha_1\sigma^2\Xi_X+\alpha_2\sigma^2\Xi_E \tag{2.37}$$

Summarizing, we arrive at

Theorem 2.1. Consider the system (2.8) with point measurement (2.9) and control law (2.22). Assume that $g(t,x,z)$ is a locally Lipschitz function satisfying $g(t,x,0)\equiv 0$ and (2.4) for a given $\sigma>0$. Let $\delta>0$, $N_0\in\mathbb{N}$ satisfy (2.14) and $N\in\mathbb{N}$ satisfy $N_0\leq N$. Let L_0 and K_0 be obtained using (2.20) and (2.21), respectively. Given $\Gamma>0$, let there exist $0<P\in\mathbb{R}^{(2N+3)\times(2N+3)}$ and scalars $\alpha_1,\alpha_2,\alpha_3>0$ such that (2.37) holds. Then, given $w(\cdot,0)\in H^1(0,1)$, the classical solution $u(t),w(x,t)$ of (2.8) subject to the control law (2.22) and the observer $\hat{w}(x,t)$ defined by (2.15)-(2.17), satisfy $u^2(t)+\|w(\cdot,t)\|_{H^1}^2+\|\hat{w}(\cdot,t)\|_{H^1}^2\leq De^{-2\delta t}\|w(\cdot,0)\|_{H^1}^2$ for $t\geq 0$ and some $D\geq 1$. Moreover, (2.37) is always feasible for N large enough and $\sigma>0$ small enough.

3. CONCLUSIONS

We studied global boundary stabilization of a semilinear heat equation under point measurement. Taking into accound the known globally Lipschitz semilinearity, we suggested a finite-dimensional nonlinear observer-based controller. Our H^1-stability analysis leads to LMIs, which are feasible for a large enough observer dimension and small enough Lipschitz constant. Our method can be extended to other semilinear PDEs. Numerical examples can be found in the recent paper (Katz and Fridman [2022]), which extends the results presented here to systems with large input delays.

REFERENCES

Christofides, P. (2001). *Nonlinear and Robust Control of PDE Systems: Methods and Applications to transport reaction processes.* Springer.

Curtain, R. (1982). Finite-dimensional compensator design for parabolic distributed systems with point sensors and boundary input. *IEEE Transactions on Automatic Control*, 27(1), 98–104.

Fridman, E. and Blighovsky, A. (2012). Robust sampled-data control of a class of semilinear parabolic systems. *Automatica*, 48, 826–836.

Kang, W. and Fridman, E. (2019). Distributed stabilization of Korteweg–de Vries–Burgers equation in the presence of input delay. *Automatica*, 100, 260–273.

Kang, W. and Fridman, E. (2020). Constrained control of 1-D parabolic PDEs using sampled in space sensing and actuation. *Systems & Control Letters*, 140, 104698.

Karafyllis, I. (2021). Lyapunov-based boundary feedback design for parabolic PDEs. *International Journal of Control*, 94(5), 1247–1260.

Karafyllis, I. and Krstic, M. (2019). Small-gain-based boundary feedback design for global exponential stabilization of one-dimensional semilinear parabolic PDEs. *SIAM Journal on Control and Optimization*, 57(3), 2016–2036.

Katz, R. and Fridman, E. (2020). Constructive method for finite-dimensional observer-based control of 1-D parabolic PDEs. *Automatica*, 122, 109285.

Katz, R. and Fridman, E. (2021a). Finite-dimensional boundary control of the linear Kuramoto-Sivashinsky equation under point measurement with guaranteed L^2-gain. *IEEE Transactions on Automatic Control.*

Katz, R. and Fridman, E. (2021b). Global stabilization of a 1D semilinear heat equation via modal decomposition and direct Lyapunov approach. Submitted.

Katz, R. and Fridman, E. (2022). Global finite-dimensional observer-based stabilization of a semilinear heat equation with large input delay. *Systems & Control Letters*, 165, 105275.

Krstic, M. and Smyshlyaev, A. (2008). *Boundary Control of PDEs: A Course on Backstepping Designs.* SIAM.

Lasiecka, I. and Triggiani, R. (2000). *Control theory for partial differential equations: Volume 1, Abstract parabolic systems: Continuous and approximation theories*, volume 1. Cambridge University Press.

Lei, J. and Khalil, H.K. (2016). High-gain-predictor-based output feedback control for time-delay nonlinear systems. *Automatica*, 71, 324–333.

Najafi, M. and Ekramian, M. (2021). Decrease the order of nonlinear predictors based on generalized-Lipschitz condition. *European Journal of Control.*

Orlov, Y., Lou, Y., and Christofides*, P.D. (2004). Robust stabilization of infinite-dimensional systems using sliding-mode output feedback control. *International Journal of Control*, 77(12), 1115–1136.

Vazquez, R. and Krstic, M. (2008). Control of 1-D parabolic PDEs with Volterra nonlinearities, part I: design. *Automatica*, 44(11), 2778–2790.

Wu, H.N., Wang, H.D., and Guo, L. (2016). Finite dimensional disturbance observer based control for nonlinear parabolic PDE systems via output feedback. *Journal of Process Control*, 48, 25–40.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Mean field games on the acceleration with state constraints ⋆

Y. Achdou * **P. Mannucci** ** **C. Marchi** *** **N. Tchou** ****

* *Université de Paris Cité and Sorbonne Université, CNRS, Laboratoire Jacques-Louis Lions, (LJLL), F-75006 Paris, France, (e-mail: achdou@ljll-univ-paris-diderot.fr)*
** *Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy, (e-mail: mannucci@math.unipd.it)*
*** *Dipartimento di Ingegneria dell'Informazione and Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Italy, (e-mail: marchi@math.unipd.it)*
**** *Univ Rennes, CNRS, IRMAR-UMR 6625, Rennes, F35000 France, (e-mail: nicoletta.tchou@univ-rennes1.fr)*

Abstract: We consider deterministic mean field games in which the agents control their acceleration and are constrained to remain in a region of $\mathbb{R}^n$. We study relaxed equilibria in the Lagrangian setting; they are described by a probability measure on trajectories. The main results of the paper concern the existence of relaxed equilibria under suitable assumptions. The fact that the optimal trajectories of the related optimal control problem solved by the agents do not form a compact set brings a difficulty in the proof of existence. The proof also requires closed graph properties of the map which associates to initial conditions the set of optimal trajectories.

Keywords: Mean field games, double integrators, Lagrangian formulation, existence of equilibria, closed graph properties
AMS: 91A13

1. INTRODUCTION

The theory of mean field games (MFGs for short) is more and more investigated since the pioneering works Lasry and Lions (2006a,b, 2007) of Lasry and Lions: it aims at studying the asymptotic behaviour of differential games (Nash equilibria) as the number of agents tends to infinity. The dynamics of the agents can be either stochastic or deterministic. Concerning the latter case, we refer to Cardaliaguet (2010) for a detailed study of deterministic MFGs in which the interactions between the agents are modeled by a nonlocal regularizing operator acting on the distribution of the states of the agents. They are described by a system of PDEs coupling a continuity equation for the density of the distribution of states (forward in time) and a Hamilton-Jacobi (HJ) equation for the optimal value of a representative agent (backward in time). If the interaction cost depends locally on the density of the distribution (hence is not regularizing), then, in the deterministic case, the available theory mostly deals with so-called variational MFGs, see Cardaliaguet et al. (2015).

⋆ YA and NT were partially supported by the ANR (Agence Nationale de la Recherche) through project ANR-16-CE40-0015-01. YA was partially supported by the chair Finance and Sustainable Development and FiME Lab (Institut Europlace de Finance). Part of the research was completed while YA was on leave at INRIA-Paris in the project Matherials. PM and CM were partially supported by GNAMPA-INdAM and by the Fondazione CaRiPaRo Project "Nonlinear Partial Differential Equations: Asymptotic Problems and Mean-Field Games".

The major part of the literature on deterministic mean field games addresses situations when the dynamics of a given agent is strongly controllable: for example, in crowd motion models, this happens if the control of a given agent is its velocity. Under the strong controllability assumption, it is possible to study realistic models in which the agents are constrained to remain in a given region K of the space state, i.e. state constrained deterministic MFGs. An important difficulty in state constrained deterministic MFGs is that nothing prevents the agents from concentrating on the boundary ∂K of the state space; let us call $m(t)$ the distribution of states at time t. Even if $m(0)$ is absolutely continuous, there may exist some $t > 0$, such that $m(t)$ has a singular part supported on ∂K and the absolute continuous part of $m(t)$ with respect to Lebesgue measure blows up near ∂K. This was first observed in some applications of MFGs to macroeconomics, see Achdou et al. (2014, 2021). From the theoretical viewpoint, the main issue is that, as we have already said, the distribution of states is generally not absolutely continuous with respect to Lebesgue measure; this makes it difficult to characterize the state distribution by means of partial differential equations. These theoretical difficulties have been addressed in Cannarsa and Capuani (2018): following ideas contained in Benamou and Brenier (2000); Benamou and Carlier (2015); Cardaliaguet et al. (2016), the authors of Cannarsa and Capuani (2018) introduce a weak or relaxed notion of equilibrium, which is defined in a Lagrangian setting rather than with PDEs. Because there may be several

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

optimal trajectories starting from a given point in the state space, the solutions of the relaxed MFG are probability measures defined on a set of admissible trajectories. Once the existence of a relaxed equilibrium is ensured, it is then possible to investigate the regularity of solutions and give a meaning to the system of PDEs and the related boundary conditions: this was done in Cannarsa et al. (2021).

On the other hand, if the agents control their acceleration instead of their velocity, the strong controllability property is lost. In Achdou et al. (2020), we have studied deterministic mean field games in the whole space $\mathbb{R}^n$ with finite time horizon T in which the dynamics of a generic agent is controlled by the acceleration. In traffic theory and also in economics, the models may require that the positions of the agents belong to a given compact subset $\overline{\Omega}$ of $\mathbb{R}^n$, and state constrained mean field games with control on the acceleration must be considered. In the present paper, we wish to investigate some examples of such mean field games and address the first step of the program followed by the authors of Cannarsa and Capuani (2018) in the strongly controllable case: we wish to prove the existence of a relaxed mean field equilibrium in the Lagrangian setting under suitable assumptions (see Definition 1 below).

The proof of existence of an equilibrium in the Lagrangian setting involves Kakutani's fixed point theorem, see Glicksberg (1952), applied to a multivalued map defined on a convex and compact set of probability measures on a suitable set of admissible trajectories (itself endowed with the $C^1([0,T];\mathbb{R}^n)\times C^0([0,T];\mathbb{R}^n)$-topology). Difficulties in applying Kakutani's fixed point theorem will arise from the fact that all the optimal trajectories do not form a compact subset of $C^1([0,T];\mathbb{R}^n)\times C^0([0,T];\mathbb{R}^n)$ (due to the lack of strong controllability). This explains why we shall need additional assumptions on the support of the initial distribution of states, see hypothesis 3 below.

2. SETTING AND MAIN RESULTS

2.1 Setting and notation

Let Ω be a bounded domain of $\mathbb{R}^n$ with a boundary $\partial\Omega$ of class C^2. For $x\in\partial\Omega$, let $n(x)$ be the unitary vector normal to $\partial\Omega$ pointing outward Ω. We will use the signed distance to $\partial\Omega$, $d:\mathbb{R}^n\to\mathbb{R}$,

$$d(x)=\begin{cases}\min\limits_{y\in\partial\Omega}|x-y|, & \text{if} \quad x\notin\Omega,\\ -\min\limits_{y\in\partial\Omega}|x-y|, & \text{if} \quad x\in\Omega.\end{cases}$$

Since $\partial\Omega$ is C^2, the function d is C^2 near $\partial\Omega$. In particular, for all $x\in\partial\Omega$, $\nabla d(x)=n(x)$.

Given a time horizon T and a pair $(x,v)\in\overline{\Omega}\times\mathbb{R}^n$, we are interested in mean field games in which the dynamics of an agent is of the form:

$$\begin{cases}\xi'(s) = \eta(s), & s\in(0,T),\\ \eta'(s) = \alpha(s), & s\in(0,T),\\ \xi(0) = x,\\ \eta(0) = v.\end{cases}\tag{1}$$

The state space is $\Xi=\overline{\Omega}\times\mathbb{R}^n$. It is convenient to define the set of admissible trajectories as follows:

$$\Gamma=\left\{\begin{array}{l}(\xi,\eta)\in C^1([0,T];\mathbb{R}^n)\times AC([0,T];\mathbb{R}^n),\\ \xi'(s)=\eta(s),\ \forall s\in[0,T],\\ (\xi(s),\eta(s))\in\Xi,\forall s\in[0,T].\end{array}\right\}.\tag{2}$$

It is a metric space with the distance $d((\xi,\eta),(\tilde{\xi},\tilde{\eta}))=\|\xi-\tilde{\xi}\|_{C^1([0,T];\mathbb{R}^n)}$.
For any $(x,v)\in\Xi$, set

$$\Gamma[x,v]=\{(\xi,\eta)\in\Gamma:\ \xi(0)=x,\ \eta(0)=v\}.\tag{3}$$

Note that $\Gamma[x,v]=\emptyset$ if $x\in\partial\Omega$ and v points outward Ω. This is the reason why we introduce $\Xi^{\rm ad}$ as follows:

$$\Xi^{\rm ad}=\{(x,v):x\in\overline{\Omega},\ v\cdot n(x)\le 0 \text{ if } x\in\partial\Omega\}\subset\Xi.\tag{4}$$

Let $\mathcal{P}(\Xi)$ be the set of probability measures on Ξ.
Let $C_b^0(\Xi;\mathbb{R})$ denote the space of bounded and continuous real valued functions defined on Ξ and let $F,G:\mathcal{P}(\Xi)\to C_b^0(\Xi;\mathbb{R})$ be bounded and continuous maps (the continuity is with respect to the narrow convergence in $\mathcal{P}(\Xi)$). Let L be a real valued, continuous and bounded from below function defined on $\Xi\times[0,T]$. Let $F[m]$ and $G[m]$ denote the images by F and G of $m\in\mathcal{P}(\Xi)$. For what follows, it is useful to introduce the positive constant $M=\sup_{(x,v,s)\in\Xi\times[0,T]}L_-(x,v,s)+\sup_{m\in\mathcal{P}(\Xi)}\|F[m]\|_{L^\infty(\Xi)}+\sup_{m\in\mathcal{P}(\Xi)}\|G[m]\|_{L^\infty(\Xi)}$. Let $\mathcal{P}(\Gamma)$ be the set of probability measures on Γ.
For $t\in[0,T]$, the evaluation map $e_t:\Gamma\to\Xi$ is defined by $e_t(\xi,\eta)=(\xi(t),\eta(t))$ for all $(\xi,\eta)\in\Gamma$.
For any $\mu\in\mathcal{P}(\Gamma)$, let the Borel probability measure $m^\mu(t)$ on Ξ be defined by $m^\mu(t)=e_t\sharp\mu$. It is possible to prove that if $\mu\in\mathcal{P}(\Gamma)$, then $t\mapsto m^\mu(t)$ is continuous from $[0,T]$ to $\mathcal{P}(\Xi)$, for the narrow convergence in $\mathcal{P}(\Xi)$. Hence, for all $(\xi,\eta)\in\Gamma$, $t\mapsto F[m^\mu(t)](\xi(t),\eta(t))$ is continuous and bounded by the constant M.

With $\mu\in\mathcal{P}(\Gamma)$ and $(\xi,\eta)\in\Gamma$, we associate the cost
$J^\mu(\xi,\eta)=$

$$\int_0^T\left(F[m^\mu(s)](\xi(s),\eta(s))+L(\xi(s),\eta(s),s)+\frac{1}{2}\left|\frac{d\eta}{dt}(s)\right|^2\right)+G[m^\mu(T)](\xi(T),\eta(T))\tag{5}$$

Definition 1. Given $m_0\in\mathcal{P}(\Xi)$, let $\mathcal{P}_{m_0}(\Gamma)$ denote the set of probability measures μ on Γ such that $e_0\sharp\mu=m_0$.
The probability measure $\mu\in\mathcal{P}_{m_0}(\Gamma)$ is a constrained mean field game equilibrium associated with the initial distribution m_0 if

$$\mathrm{supp}(\mu)\subset\bigcup_{(x,v)\in\mathrm{supp}(m_0)}\Gamma^{\mu,\mathrm{opt}}[x,v].\tag{6}$$

where

$$\Gamma^{\mu,\mathrm{opt}}[x,v]=\left\{(\xi,\eta)\in\Gamma[x,v]\ :\ J^\mu(\xi,\eta)=\min_{(\widetilde{\xi},\widetilde{\eta})\in\Gamma[x,v]}J^\mu(\widetilde{\xi},\widetilde{\eta})\right\},$$

2.2 Closed graph properties and bounds related to optimal trajectories

In all what follows, L,F and G satisfy the assumptions made in §2.1. An important step in the proof of existence of constrained mean field game equilibria is the closed graph property:

Proposition 1. Consider a closed subset Θ of $\Xi^{\rm ad}$. Given $\mu\in\mathcal{P}(\Xi)$, assume that for all sequences $(x^i,v^i)_{i\in\mathbb{N}}$ such that for all $i\in\mathbb{N}$, $(x^i,v^i)\in\Theta$ and $\lim\limits_{i\to+\infty}(x^i,v^i)=(x,v)\in\Theta$, the following holds: if $x\in\partial\Omega$, then

$$\left((v^i \cdot \nabla d(x^i))_+\right)^3 = o\left(\left|d(x^i)\right|\right), \tag{7}$$

(note that (7) is meaningful for i large enough, because d is C^1 near $\partial\Omega$); then the graph of the multivalued map

$$\Gamma^{\rm opt} : \Theta \rightrightarrows \Gamma,$$
$$(x,v) \mapsto \Gamma^{\mu,{\rm opt}}[x,v]$$

is closed, which means: for any sequence $(y^i,w^i)_{i\in\mathbb{N}}$ such that for all $i\in\mathbb{N}$, $(y^i,w^i)\in\Theta$ with $(y^i,w^i)\to(y,w)$ as $i\to\infty$, consider a sequence $(\xi^i,\eta^i)_{i\in\mathbb{N}}$ such that for all $i\in\mathbb{N}$, $(\xi^i,\eta^i)\in\Gamma^{\rm opt}[y^i,w^i]$; if (ξ^i,η^i) tends to (ξ,η) uniformly, then $(\xi,\eta)\in\Gamma^{\mu,{\rm opt}}[y,w]$.

Definition 2. For numbers r and C, let us set

$$K_r = \{(x,v)\in\Xi :\ |v|\le r\}, \tag{8}$$

$$\Gamma_C = \left\{ (\xi,\eta)\in\Gamma : \left| \begin{array}{l} (\xi(t),\eta(t))\in K_C, \quad \forall t\in[0,T], \\ \left\| \dfrac{d\eta}{dt} \right\|_{L^2(0,T;\mathbb{R}^n)} \le C \end{array} \right. \right\} \tag{9}$$

The set Γ_C is a compact subset of Γ.

Proposition 2. Given $r>0$, let us define

$$\Theta_r = \Theta \cap K_r, \tag{10}$$

where K_r is defined by (8) and Θ is a closed subset of $\Xi^{\rm ad}$ which satisfies the assumption in Proposition 1.
The value function

$$u^\mu(x,v) = \inf_{(\xi,\eta)\in\Gamma[x,v]} J^\mu(\xi,\eta,\eta') \tag{11}$$

is continuous on Θ_r and there exists a positive number $C=C(r,M)$ independent of $\mu\in\mathcal{P}(\Xi)$ such that if $(x,v)\in\Theta_r$, then $\Gamma^{\mu,{\rm opt}}[x,v]\subset\Gamma_C$.

Hypothesis 3. There exists a positive number r such that the initial distribution of states is a probability measure m_0 on Ξ supported in Θ_r, where Θ_r is a closed subset of $\Xi^{\rm ad}$ as in (10).

Let $C=C(r,M)$ be the constant appearing in Proposition 2 (uniform w.r.t. μ), and Γ_C be the compact subset of Γ defined by (9); clearly, Γ_C is a Radon metric space. From Prokhorov theorem, see (Ambrosio et al., 2005, Theorem 5.1.3), the set $\mathcal{P}(\Gamma_C)$ is compact for the narrow convergence of measures.

As above, $\mathcal{P}_{m_0}(\Gamma_C))$ denotes the set of probability measures μ on Γ_C such that $e_0\sharp\mu = m_0$.

Remark 1. Note that Γ_C (endowed with the metric of the $C^1\times C^0$-convergence of (ξ,η)) is a Polish space (because it is compact). Using Kuratowski and Ryll-Nardzewski theorem, Kuratowski and Ryll-Nardzewski (1965), it can be proved that there exists a measurable selection $j:\Theta_r\to\Gamma_C$. Then $j\sharp m_0$ belongs to $\mathcal{P}_{m_0}(\Gamma_C)$. The set $\mathcal{P}_{m_0}(\Gamma_C)$ is not empty.

Standard arguments from the calculus of variations yield that for each $\mu\in\mathcal{P}_{m_0}(\Gamma_C)$ and $(x,v)\in\Xi^{\rm ad}$, $\Gamma^{\mu,{\rm opt}}[x,v]$ is not empty. Moreover, from Proposition 2, $\Gamma^{\mu,{\rm opt}}[x,v]\subset\Gamma_C$ for all $(x,v)\in\Theta_r$.

Proposition 4. Under Hypothesis 3, let $C=C(r,M)$ be chosen as in Proposition 2.
Let a sequence of probability measures $(\mu_i)_{i\in\mathbb{N}}$, $\mu_i\in\mathcal{P}_{m_0}(\Gamma_C)$, be narrowly convergent to $\mu\in\mathcal{P}(\Gamma_C)$. Let $(x^i,v^i)_{i\in\mathbb{N}}$ be a sequence with $(x^i,v^i)\in\Theta_r$ which converges to (x,v). Consider a sequence $(\xi^i,\eta^i)_{i\in\mathbb{N}}$ such that for all $i\in\mathbb{N}$, $(\xi^i,\eta^i)\in\Gamma^{\mu_i,{\rm opt}}[x^i,v^i]$. If $(\xi^i,\eta^i)_{i\in\mathbb{N}}$ tends to (ξ,η) uniformly, then $(\xi,\eta)\in\Gamma^{\mu,{\rm opt}}[x,v]$. In other words, the multivalued map $(x,v,\mu)\mapsto\Gamma^{\mu,{\rm opt}}[x,v]$ has closed graph.

2.3 Existence of a mean field game equilibrium

Theorem 5. Under the assumptions made on L, F and G in §2.1 and Hypothesis 3, let $C=C(r,M)$ be chosen as in Proposition 2. There exists a constrained mean field game equilibrium $\mu\in\mathcal{P}_{m_0}(\Gamma_C)$, see Definition 1. Moreover, $t\mapsto e_t\sharp\mu\in C^{1/2}([0,T];\mathcal{P}(K_C))$, ($K_C$ is defined in (8) and $\mathcal{P}(K_C)$ is endowed with the Kantorovitch-Rubinstein distance).

The proof of Theorem 5 is inspired from that of Cannarsa and Capuani in Cannarsa and Capuani (2018). It consists of applying Kakutani's fixed point theorem to the multivalued map E from $\mathcal{P}_{m_0}(\Gamma_C)$ to $\mathcal{P}_{m_0}(\Gamma_C)$ as follows: for any $\mu\in\mathcal{P}_{m_0}(\Gamma_C)$,

$$E(\mu) = \left\{ \hat\mu\in\mathcal{P}_{m_0}(\Gamma_C)\ : \ {\rm supp}(\hat\mu_{(x,v)})\subset\Gamma^{\mu,{\rm opt}}[x,v] \text{ for } m_0\text{-a-a } (x,v)\in\Xi \right\},$$

where $(\hat\mu_{(x,v)})_{(x,v)\in\Xi}$ is the m_0-almost everywhere uniquely defined Borel measurable family of probability measures which disintegrates $\hat\mu$. For that, a key step is Proposition 4.

Remark 2. In dimension one and for a running cost quadratic in α, it is possible to obtain refined results under a slightly stronger assumption on the running cost, namely that it does not favor the trajectories which exit the domain. In particular, the closed graph property can be proved to hold on the whole set $\Xi^{\rm ad}$, and concerning mean field games, no assumptions are needed on the support of m_0 by contrast with Theorem 5.

Definition 3. A pair (u,m), where u is a measurable function defined on $\Xi\times[0,T]$ and $m\in C^0([0,T];\mathcal{P}(\Xi))$, is called *a mild solution of the mean field game*, if there exists a constrained mean fied game equilibrium μ for m_0 (see Definition 1) such that:

i) $m(t)=e_t\sharp\mu$;
ii) $\forall(x,v)\in\Xi^{\rm ad}$, $u(x,v,t)$ is given by

$$u(x,v,t) = \inf_{(\xi,\eta)\in\Gamma[x,v,t]} J^\mu(t,\xi,\eta)$$

where $\Gamma[x,v,t]$ is the set of admissible trajectories starting from (x,v) at $s=t$ and

$$J^\mu(t,\xi,\eta) = \int_t^T F[m(s)](\xi(s),\eta(s)) + L(\xi(s),\eta(s),s) + \frac{1}{2}|\frac{d\eta}{ds}(s)|^2 + G[m(T)](\xi(T),\eta(T)).$$

Remark 3. It is tempting to say that a mild solution (u,m) is a very weak solution of a boundary value problem related to a system of PDEs posed in $\Omega\times[0,T]$, composed of a Hamilton-Jacobi equation for finding the optimal strategies, and of a Fokker-Planck equation for the evolution of m. However, this system should be supplemented with boundary conditions on $\partial\Omega\times(0,T)$. This aspect is particularly tricky because u blows up on some part of the boundary.

A corollary of Theorem 5 is:

Corollary 1. Under the assumptions of Theorem 5, there exists a mild solution (u, m).

Moreover, $m \in C^{\frac{1}{2}}([0,T];\mathcal{P}(K_C))$.

Remark 4. Under classical monotonicity assumptions for F and G, see e.g. Cannarsa and Capuani (2018), the mild solution is unique.

REFERENCES

Achdou, Y., Buera, F., Lasry, J.M., Lions, P.L., and Moll, B. (2014). Partial differential equation models in macroeconomics. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 372(2028), 20130397, 19. doi:10.1098/rsta.2013.0397. URL http://dx.doi.org/10.1098/rsta.2013.0397.

Achdou, Y., Han, J., Lasry, J.M., Lions, P.L., and Moll, B. (2021). Income and wealth distribution in macroeconomics: A continuous-time approach. *The review of economic studies*.

Achdou, Y., Mannucci, P., Marchi, C., and Tchou, N. (2020). Deterministic mean field games with control on the acceleration. *NoDEA Nonlinear Differential Equations Appl.*, 27(3), Paper No. 33. doi:10.1007/s00030-020-00634-y. URL https://doi.org/10.1007/s00030-020-00634-y.

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel.

Benamou, J.D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3), 375–393. doi:10.1007/s002110050002. URL http://dx.doi.org/10.1007/s002110050002.

Benamou, J.D. and Carlier, G. (2015). Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1), 1–26. doi:10.1007/s10957-015-0725-9. URL http://dx.doi.org/10.1007/s10957-015-0725-9.

Cannarsa, P. and Capuani, R. (2018). Existence and uniqueness for mean field games with state constraints. In *PDE models for multi-agent phenomena*, volume 28 of *Springer INdAM Ser.*, 49–71. Springer, Cham.

Cannarsa, P., Capuani, R., and Cardaliaguet, P. (2021). Mean field games with state constraints: from mild to pointwise solutions of the PDE system. *Calc. Var. Partial Differential Equations*, 60(3), Paper No. 108, 33. doi:10.1007/s00526-021-01936-4. URL https://doi.org/10.1007/s00526-021-01936-4.

Cardaliaguet, P. (2010). Notes on mean field games. Preprint, 2011.

Cardaliaguet, P., Graber, J., Porretta, A., and Tonon, D. (2015). Second order mean field games with degenerate diffusion and local coupling. *NoDEA Nonlinear Differential Equations Appl.*, 22(5), 1287–1317.

Cardaliaguet, P., Mészáros, A.R., and Santambrogio, F. (2016). First order mean field games with density constraints: pressure equals price. *SIAM J. Control Optim.*, 54(5), 2672–2709. doi:10.1137/15M1029849. URL https://doi.org/10.1137/15M1029849.

Glicksberg, I.L. (1952). A further generalization of the Kakutani fixed theorem, with application to Nash equilibrium points. *Proc. Amer. Math. Soc.*, 3, 170–174. doi:10.2307/2032478. URL https://doi.org/10.2307/2032478.

Kuratowski, K. and Ryll-Nardzewski, C. (1965). A general theorem on selectors. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.*, 13, 397–403.

Lasry, J.M. and Lions, P.L. (2006a). Jeux à champ moyen. I. Le cas stationnaire. *C. R. Math. Acad. Sci. Paris*, 343(9), 619–625.

Lasry, J.M. and Lions, P.L. (2006b). Jeux à champ moyen. II. Horizon fini et contrôle optimal. *C. R. Math. Acad. Sci. Paris*, 343(10), 679–684.

Lasry, J.M. and Lions, P.L. (2007). Mean field games. *Jpn. J. Math.*, 2(1), 229–260.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Data-driven coarse-graining of agent-based models through stochastic differential equations ⋆

Asima Azmat * **Kaili Wang** ** **Felix Dietrich** **

* *Siemens AG, Munich, Germany.*
** *Technical University of Munich, Department of Informatics, Munich, Germany (corr. e-mail: felix.dietrich@tum.de).*

Abstract: Macroscopic, coarse descriptions of microscopic, agent-based dynamical systems are useful for tasks such as optimization, bifurcation analysis, and control. Once suitable coarse variables are defined, their dynamics can be either derived analytically or approximated in a data-driven fashion. For many agent-based systems, this coarse-graining procedure requires appropriate closure terms or stochastic elements on the macroscopic scale to summarize degrees of freedom of the agents. In this contribution, we identify effective stochastic differential equations (SDE) for coarse observables of agent-based simulations. These SDE then act as surrogate models on the macroscopic scale. We approximate the drift and diffusivity functions for these SDE through neural networks. Based on earlier work, the loss function is inspired by the structure of established stochastic numerical integrators, in particular Euler-Maruyama and Milstein schemes. We consider cases where the coarse collective observables are known in advance, and where they must be found with data-driven methods. We demonstrate the feasibility on data from an egress simulation of pedestrians in two-dimensional continuous space (with the crowd simulation software Vadere).

Keywords: Stochastic system identification, Simulation of stochastic systems, Statistical data analysis, Software for system identification, Machine learning for environmental applications

1. INTRODUCTION

Agent-based simulations provide accurate and detailed insights into dynamical systems. However, the simulation time and computational burden of these microscopic simulations often prohibit large-scale parameter studies or overviews on coarser scales. These tasks can be done with newly devised coarser models, but those often suffer from low accuracy compared to the agent-based approach. As a remedy, we aim to learn coarse models directly from the agent-based data. We identify effective stochastic differential equations (SDE) for coarse observables of fine-grained particle- or agent-based simulations. These SDE then provide coarse surrogate models of the fine scale dynamics. The coarse variables in question for this contribution are the infection states for a viral disease that is spreading through a local population. The crowd simulation software Vadere (`www.vadere.org`) is used for the agent-based modelling and simulation. To learn the SDE, we approximate the drift and diffusivity functions of the effective SDE through neural networks, which can be thought of as effective stochastic ResNets. The loss function is inspired by the structure of established stochastic numerical integrators (here, the Euler-Maruyama integrator; more intricate Milstein or Runge-Kutta integrators are considered in our paper, Dietrich et al. (2021)). The SDE approximations can benefit from error analysis of these underlying numerical schemes. They also lend themselves naturally to "physics-informed" gray-box identification when approximate coarse models, such as mean field equations, are available. The learning procedure we use does not require long trajectories, works on scattered snapshot data, and is designed to naturally handle different time steps per snapshot.

In related work, residual networks (ResNets He et al. (2016), but see also Rico-Martínez et al. (1992)) successfully employ a forward-Euler integrator based approach to create very deep architectures. This has inspired followup work to include other integrator schemes for deterministic, ordinary differential equations (ODE) such as symplectic integrators Bertalan et al. (2019); Zhu et al. (2020). From ResNets (He et al., 2016) and Neural ODEs (Chen et al., 2018; Rico-Martínez et al., 1992) to DeepONets (Lu et al., 2021), the identification of dynamical systems from (discrete) spatio-temporal data is a booming business because it is now comparatively easy to program and train neural networks. The time-honored SDE estimation techniques for local drift and diffusivity can now be synthesized in a (more or less global) surrogate model.

The software is available in the following repository: `https://gitlab.com/felix.dietrich/sde-identification`.

2. MATHEMATICAL PROBLEM SETTING

In this section, we follow the setting from our recent paper (Dietrich et al., 2021). We discuss the SDE

⋆ F.D. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 468830823.

$$\mathrm{d}x_t = f(x_t)\mathrm{d}t + \sigma(x_t)\mathrm{d}W_t, \tag{1}$$

where $f : R^n \to R^n$ and $\sigma : R^n \to R^{n\times n}$ are smooth, possibly nonlinear functions; every $\sigma(x)$ is positive and bounded away from zero (and a positive-definite matrix if $n > 1$), and W_t a Wiener process such that for $t > s$, $W_t - W_s \sim \mathcal{N}(0, t-s)$. We assume we have access to a set of N snapshots $D = \{(x_1^{(k)}, x_0^{(k)}, h^{(k)})\}_{k=1}^N$, where $x_0^{(k)}$ are points scattered in the state space of (1) and the value of $x_1^{(k)}$ results from the evolution of Eq. 1 under a small time-step $h^{(k)} > 0$, starting at $x_0^{(k)}$. The snapshots are samples from a distribution, $x_0 \sim p_0$, and the transition densities $x_1 \sim p_1(\cdot|x_0, h)$ are associated with (1) for a given time-step $h > 0$, which in turn is chosen from some distribution p_h. The joint data-generating distribution is therefore given by $p(x_0, x_1, h) = p_1(x_1|x_1, h)p_0(x_0)p_h(h)$. Alternatively, the data could be collected along a long trajectory $\{x_{t_i}\}$ of (1) with sample frequency $h_i > 0$, that is, $t_{i+1} = t_i + h_i$. The problem is to identify the drift f and diffusivity σ through two neural networks $f_\theta : R^n \to R^n$ and $\sigma_\theta : R^n \to R^{n\times n}$, parameterized by their weights θ, only from the data in D. We assume the points in D are sampled sufficiently densely in the region of interest.

2.1 Identification with the Euler-Maruyama scheme

We now formulate and rationalize the loss term that we use to train the neural networks f_θ and σ_θ. The Euler-Maruyama scheme is a simple method to integrate Eq. 1 over a small time $h > 0$:

$$x_1 = x_0 + hf(x_0) + \sigma(x_0)\,\delta W_0, \tag{2}$$

where $h > 0$ is small and δW_0 is a vector of n random variables, all i.i.d. and normally distributed around zero with variance h. The convergence of Eq. 2 for $h \to 0$ has been studied at length; We refer to standard literature (Pavliotis, 2014). We can use this idea to construct a loss function for training the networks σ_θ and f_θ simultaneously. We initially restrict the discussion to the case $n = 1$ for simplicity. Essentially, conditioned on x_0 and h, we can think of x_1 as a point drawn from a multivariate normal distribution

$$x_1 \sim \mathcal{N}\left(x_0 + hf(x_0), h\sigma(x_0)^2\right). \tag{3}$$

In the training data set D, we only have access to triples $(x_0^{(k)}, x_1^{(k)}, h^{(k)})$, and not the drift f and diffusivity σ. To approximate them, we define the probability density p_θ of the normal distribution Eq. 3 and then, given the neural networks f_θ and σ_θ, ask that the log-likelihood of the data D under the assumption in equation Eq. 3 is high:

$$\begin{aligned}\theta := arg\max_{\hat\theta} E\left[\log p_{\hat\theta}\left(x_1|x_0,h\right)\right] \approx \\ arg\max_{\hat\theta} \frac{1}{N}\sum_{k=1}^{N} \log p_{\hat\theta}\left(x_1^{(k)}\,\middle|\,x_0^{(k)}, h^{(k)}\right).\end{aligned} \tag{4}$$

We can now formulate the loss function that will be minimized to obtain the neural network weights θ. The logarithm of the well-known probability density function of the normal distribution, together with the mean and variance from Eq. 3, yields the loss to minimize during training,

$$\begin{aligned}&\mathcal{L}(\theta|x_0, x_1, h) := \\ &\frac{\left(x_1 - x_0 - hf_\theta(x_0)\right)^2}{h\,\sigma_\theta(x_0)^2} + \log\left|h\sigma_\theta\left(x_0\right)^2\right| + \log(2\pi).\end{aligned} \tag{5}$$

This formula can easily be generalized to higher dimensions, and we use such generalizations for examples in more than one dimension. Minimizing $\mathcal{L}$ in Eq. 5 over the data D implies maximization of the log marginal likelihood Eq. 4 with the constant terms removed (as they do not influence the minimization) (Pavliotis, 2014). Likelihood estimation in combination with the normal distribution is used in many variational and generative approaches (Goodfellow et al., 2014; Kingma and Welling, 2014; Li et al., 2020; Yildiz et al., 2018; Yang et al., 2021). Note that here, the step size $h^{(k)}$ is defined *per snapshot*, so it is possible that it has different values for every index k. This is especially useful in simulations where the time step is determined as part of the scheme, e.g. a Gillespie simulation.

3. COMPUTATIONAL EXPERIMENTS

We tested our implementation with examples in one dimension first, with drift f and diffusivity σ defined through $f(x_t) = -2x_t^3 - 4x_t + 1.5$, $\sigma(x_t) = 0.05x_t + 0.5$. A comparison between the learned and true functions is shown in Fig. 1, along with densities obtained from sampling the true and approximate SDE. We trained networks with both Euler-Maruyama and Milstein loss functions, with very comparable results—even the training and validation curves were very similar. When increasing the dimension from $n = 1$ to $n = 20$, approximation quality decreases after $n = 6$.

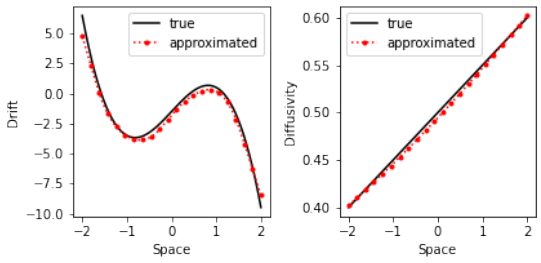


Fig. 1. Approximation vs. true values of f and σ in the first example.

In Fig. 2, we demonstrate that our approach can learn non-diagonal diffusivity matrices. We sample initial points $x_0 \in [-.3, .3]^3$, randomly sample a lower-triagonal diffusivity matrix Σ with positive eigenvalues, which we use as the (constant) diffusivity $\sigma(x) = \Sigma$, and set the drift to be $f(x) = -x$. The absolute error between original and (the average over) identified matrix entries is smaller than 0.01, and the standard deviation of the diffusivity values over the entire data set is smaller than 0.003 for all elements of the matrix. Note that the network σ_θ is still a nonlinear function over the state space, not a constant. To ensure the matrix has positive eigenvalues, we pass the real-valued output of the neurons that encode the diagonal of the matrix through a soft-plus activation function.

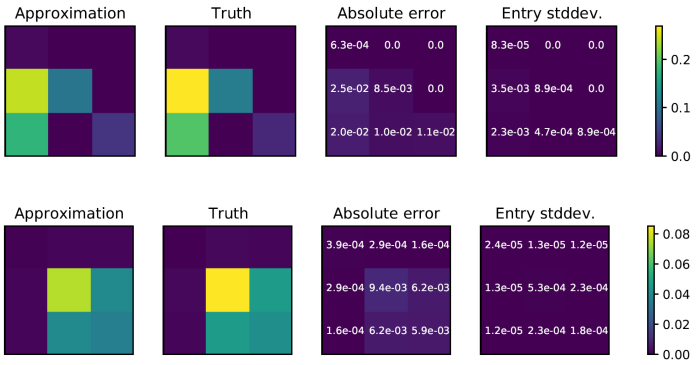


Fig. 2. Lower-triangular (top row) and full symmetric, positive definite (bottom row) diffusivity matrix approximation. The panels show the averaged network output over the uniformly sampled data inside the training region, the true diffusivity matrices, the absolute error between the results shown in the first and second columns, and the deviation of the network matrix output over the training region. The color range is the same for all four plots.

3.1 Toy example: increasing dimension from 1D to 20D

We define $f(x) = -(4x^3 - 8x + 3)/2$ and $\sigma(x) = 5e - 2x + 0.5$, where all operations are meant coordinate-wise (e.g. x^3 computes the third power of each individual coordinate of the vector). Figure 3 (left) illustrates how the training and validation losses change when increasing n from 1 to 20. The loss (Eq. 5) is adapted to the changing dimensionality by re-adding the constant term $\log(2\pi)n$. The increase in loss can be explained by the constant number of points ($N = 10,000$) we used: Increasing the intrinsic dimension of the problem by sampling the input data in the n-dimensional cube $[-2, 2]^n$ causes the data sampling to get sparser and sparser. By increasing the number of training data points linearly with the dimension (while keeping the number of training iterations per dimension constant), we can see that the training loss is relatively small even for $n = 12$ (Fig. 3, right).

3.2 Coarse-graining an agent-based infection model

We now demonstrate that the SDE learning approach can be used to effectively coarse-grain a crowd simulation model. The crowd simulation was performed using the software Vadere (www.vadere.org). Fig. 4 (left) shows our test scenario. Agents are placed in a square box of $30m \times 30m$, and can infect neighbors in a radius of $1.5m$ with a probability of 0.1% every simulated second. The other two panels show scenarios that we are currently investigating: a bottleneck scenario with dynamically changing crowd positions, and a classroom with students moving in and out of the room over the course of multiple days. Fig. 5 shows sample paths of the learned SDE compared to test trajectories sampled with the simulation software.

4. CONCLUSIONS

Training neural networks with loss functions based on numerical integrators such as Euler-Maruyama and Milstein has several limitations. If the time step between many

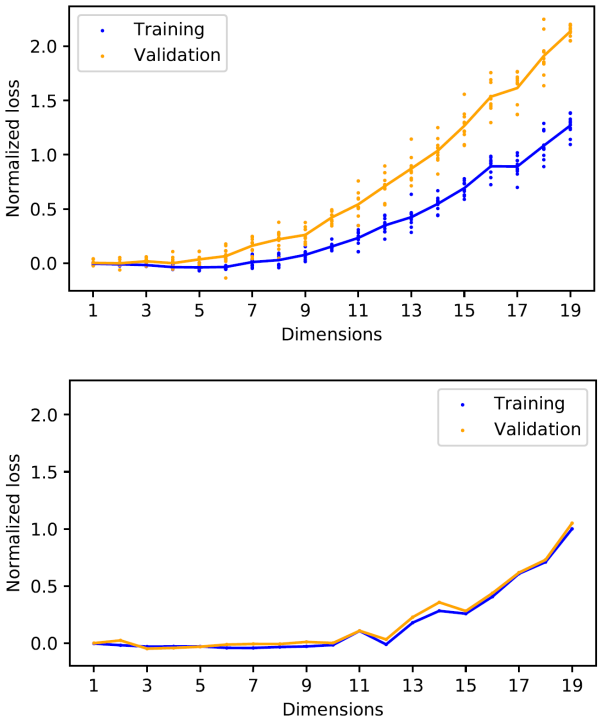


Fig. 3. Normalized training and validation loss when increasing the intrinsic dimension of the problem from $n = 1$ to $n = 20$. The top panel shows results when keeping the number of samples constant at $N = 10,000$, the bottom panel shows what happens when increasing $N = 10,000 \times n$, but training for the same number of iterations per dimension ($Epochs = 1000/n$).

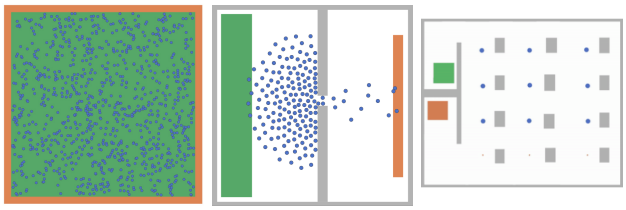


Fig. 4. Crowd scenarios simulated in Vadere. Left; A crowd with 1000 agents is placed in a square that mimics a concert setting. Center: 100 agents move through a bottleneck towards a target. Right: a classroom scenario with students moving in and out of the room over the course of multiple days. For each scenario, we perform 10,000 short simulation runs, and between 0 and 70% of the crowd is infected with an air-borne disease that can spread to neighbors.

samples is too small, we cannot accurately identify the drift, because the diffusivity term will dominate. This could be mitigated by starting with small h, estimating the diffusivity, and then estimating the drift with sub-sampled trajectories. Even with infinite data the drift is difficult to estimate, because the time step also has to go to zero (Pavliotis, 2014). Conversely, if the time step is too big, we cannot accurately identify the diffusivity. If the dynamics of the coarse-grained observables include rare events, then learning the corresponding SDE is a challenge, because these events will not be present in a

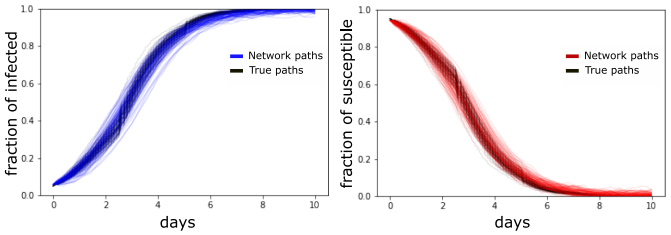


Fig. 5. Sample paths drawn from the learned SDE, compared to measured data, in the concert scenario.

lot of snapshots, and hence not a lot of data is available to learn them. Using a loss function based on Lévy noise SDE integrators might provide a useful surrogate for SDEs with rare-events.

There are many possible extensions and applications of our work. An important task is to find explicit probability density functions for integrators using other noise types, such as Lévy or Poisson noise. Analyzing integrators based on other types of calculus such as Heun's (Burrage et al., 2004) method may help impose more variegated types of priors during training. Using an adjoint method (Liu et al., 2020) can allow us to propagate the loss gradients back through longer time series.

In terms of applications, many more particle- and agent-based models can be coarse grained. For many of these, it is important to find coarse variables, and latent space techniques such as VAE-type architectures and Diffusion Maps (Coifman and Lafon, 2006) will help to identify them. These techniques can help to construct local SDE models in the process of larger-scale simulations to guide sampling and exploration. Another clear next step based on numerical analysis is the identification of stochastic partial differential equations (SPDEs).

ACKNOWLEDGEMENTS

This extended abstract is based on our work with several collaborators: Alexei Makeev, George Kevrekidis, Nikolaos Evangelou, Tom Bertalan, Sebastian Reich, and Ioannis G. Kevrekidis (see Dietrich et al. (2021)). We also discussed the crowd simulation scenarios and infectious disease model with Gerta Köster and Simon Rahn. F.D. was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 468830823.

REFERENCES

Bertalan, T., Dietrich, F., Mezić, I., and Kevrekidis, I.G. (2019). On learning hamiltonian systems from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12), 121107. doi:10.1063/1.5128231.

Burrage, K., Burrage, P.M., and Tian, T. (2004). Numerical methods for strong solutions of stochastic differential equations: an overview. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460(2041), 373–402. doi:10.1098/rspa.2003.1247.

Chen, R.T.Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations. In *NeurIPS conference 2018*.

Coifman, R.R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), 5–30. doi:10.1016/j.acha.2006.04.006.

Dietrich, F., Makeev, A., Kevrekidis, G., Evangelou, N., Bertalan, T., Reich, S., and Kevrekidis, I.G. (2021). Learning effective stochastic differential equations from microscopic simulations: combining stochastic numerics and deep learning. *arXiv:2106.09004*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, 2672–2680. Curran Associates, Inc.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/cvpr.2016.90.

Kingma, D.P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Li, X., Wong, T.K.L., Chen, R.T., and Duvenaud, D. (2020). Scalable gradients for stochastic differential equations. *arXiv:2001.01328*.

Liu, J., Long, Z., Wang, R., Sun, J., and Dong, B. (2020). Rode-net: Learning ordinary differential equations with randomness from data. *arXiv:2006.02377*.

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G.E. (2021). Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3), 218–229. doi:10.1038/s42256-021-00302-5.

Pavliotis, G.A. (2014). *Stochastic Processes and Applications*. Springer New York. doi:10.1007/978-1-4939-1323-7.

Rico-Martínez, R., Krischer, K., Kevrekidis, I., Kube, M., and Hudson, J. (1992). Discrete- vs. continuous-time nonlinear signal processing of Cu electrodissolution data. *Chemical Engineering Communications*, 118(1), 25–48. doi:10.1080/00986449208936084.

Yang, S., Wong, S.W.K., and Kou, S.C. (2021). Inference of dynamic systems from noisy and sparse data via manifold-constrained gaussian processes. *Proceedings of the National Academy of Sciences*, 118(15), e2020397118. doi:10.1073/pnas.2020397118.

Yildiz, C., Heinonen, M., Intosalmi, J., Mannerstrom, H., and Lahdesmaki, H. (2018). Learning stochastic differential equations with with Gaussian Processes without gradient matching. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. doi:10.1109/mlsp.2018.8516991.

Zhu, A., Jin, P., and Tang, Y. (2020). Deep hamiltonian networks based on symplectic integrators. *arXiv:2004.13830*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Dead-Time Compensation as an Observer-Based Design

Leonid Mirkin * **Dennis Zanutto** **

* *Faculty of Mechanical Eng., Technion—IIT, Haifa 3200003, Israel*
(e-mail: mirkin@technion.ac.il)
** *DEIB, Politecnico di Milano, 20133 Milan, Italy*
(e-mail: dennis.zanutto@mail.polimi.it)

Abstract: In this note we study discrete-time dead-time compensation from the viewpoint of the observer-based design procedure. We show that the discrete equivalent of the observer-predictor architecture can be derived *ab initio* via classical state-feedback and observer arguments under mild assumptions. The resulting observer is reduced order and we show that this choice is justifiable even if corresponding state measurement channels are noisy.

Keywords: Time-delay systems, observer-based control, H_2 filtering, reduced-order observers.

1. INTRODUCTION

Since its introduction by Otto J. M. Smith (1957), the notion of delay compensation (or dead-time compensation, DTC) plays an important role in the control of time-delay systems. The use of DTC configurations makes it possible to reduce various delay problems to their delay-free equivalents. Initially, the idea was viewed as an ingenious transformation, simplifying the stability analysis, see (Manitius and Olbrot, 1979; Artstein, 1982; Furukawa and Shimemura, 1983; Fiagbedzi and Pearson, 1990) among other results. Later on, DTC was shown to be an intrinsic part of stabilizing and optimal H_2 (Mirkin and Raskin, 2003), H_∞ (Meinsma and Zwart, 2000; Mirkin, 2003), and L_1 (Mirkin, 2006) controllers in the single I/O delay case, as well as in some multiple I/O delay settings (Meinsma and Mirkin, 2005; Mirkin et al., 2011).

Historically, the emphasis in the DTC research was laid on continuous-time systems. A possible reason is that the delay element is infinite dimensional in continuous time. Hence, the reduction to a finite-dimensional delay-free case is particularly appealing there. Discrete delays are finite dimensional themselves, so the gain in compensating them is less apparent. Moreover, discrete formulae are often bulkier, which is especially hindering in optimization problems. Still, the finite-dimensional nature of discrete delays offers an opportunity to understand the rationale behind DTC via accessible means.

Our first goal in this note is to *derive* a discrete version of the observer-predictor DTC architecture of (Furukawa and Shimemura, 1983) via classical observer-based arguments, from scratch. We show that this derivation involves only two choices, which are justified. The first one is to keep deadbeat modes of the plant, originated in the delay dynamics, untouched in the state-feedback phase of the design. The second choice is to use a reduced-order observer to estimate the state. This is justified by the fact that the history of control inputs, which is a part of the plant state in the delay case, can be expected to be known accurately. We believe that the proposed derivation offers a new insight into the well-studied observer-predictor architecture.

The second, and most important, goal of this note is to demonstrate that the reduced-order observer architecture is *justified* in this case even if the control input itself is noisy. This conclusion may appear counterintuitive. Reduced-order observers are known to be sensitive to measurement noise (Anderson and Moore, 1989). This property was even used in (Krstic, 2009, Sec. 3.3) to justify the adoption of full-order alternatives in continuous-time DTC. We study this issue by posing the observer design problem as an H_2 optimization under noisy measurements. Although the problem is somewhat unorthodox, we end up with its transparent analytic solution, which has a reduced-order observer form regardless the noise intensity. We are not aware of any other problem, where the steady-state Kalman filter is of a reduced-order form under nonsingular measurement noise. This way we show that the single-delay DTC is a rare case where reduced-order observers can be vindicated of their "original guilt."

This is a conference version of (Mirkin and Zanutto, 2022), to which a reader is referred for all proofs.

Notation By $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ we indicate the sets of real, integer, and natural (positive integer) numbers, respectively, and $\mathbb{Z}_{i_1..i_2} := \{i \in \mathbb{Z} \mid i_1 \le i \le i_2\}$. The complex-conjugate transpose of a matrix M is denoted by M' and its Frobenius norm $\|M\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(M'M)}$. If a matrix M is square, $\mathrm{spec}(M)$ stands for its spectrum, i.e. the (multi) set of all its eigenvalues. By $\{a\}^n$ we understand the multiset containing n copies of a.

We say that a discrete-time linear shift-invariant system G is an H_2 system if it is causal and its impulse response matrix is square summable. If $G \in H_2$, then its H_2-norm

$$\|G\|_2 = \Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} \|G(\mathrm{e}^{\mathrm{j}\theta})\|_{\mathrm{F}}^2 \mathrm{d}\theta\Big)^{1/2},$$

where $G(z)$ is the transfer function of G. We use the compact notation

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] := D + C(zI - A)^{-1}B$$

for transfer functions in terms of their state-space realizations.

⋆ Supported by the Israel Science Foundation (grant no. 3177/21) and, in part, by Sakranut Graydah at Politecnico di Milano.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2. DTC VIA OBSERVER-BASED DESIGN

Consider an input-delay discrete system described by

$$\begin{cases} x[t+1] = Ax[t] + Bu[t-\tau] \\ y[t] = Cx[t] + Du[t-\tau] \end{cases}, \tag{1}$$

where $u[t] \in \mathbb{R}^m$ is a control input, $y[t] \in \mathbb{R}^p$ is a measured output, $x[t] \in \mathbb{R}^n$ is an internal signal, and $\tau \in \mathbb{N}$ is a delay. We assume hereafter that

$\boldsymbol{\mathcal{A}_1}$: (A, B) is stabilizable and (C, A) is detectable,

which is necessary for stabilization. We also assume that

$\boldsymbol{\mathcal{A}_2}$: (A, B) has no unreachable modes at the origin,

which is required only to simplify the exposition, the results remain qualitatively the same if this assumption is omitted.

The vector x is not a state of (1). Indeed, by the state we understand a history accumulator, whose knowledge at any time instant $t = t_0$ is the only information about the past required to calculate the behavior of the system for $t > t_0$. It is readily seen that the knowledge of $x[t_0]$ in (1) should be complemented by that of $u[t]$ for all $t \in \mathbb{Z}_{t_0-\tau..t_0-1}$. A state of (1) may then be chosen as

$$x_\tau[t] := \begin{bmatrix} x[t] \\ u[t-\tau] \\ \vdots \\ u[t-1] \end{bmatrix} \in \mathbb{R}^{n+m\tau}. \tag{2}$$

With this choice, the measured subset of the state changes as well, because the history of the control signal may be safely assumed to be measurable. Hence, the measured variable is

$$y_\tau[t] := \begin{bmatrix} y[t] \\ u[t-\tau] \\ \vdots \\ u[t-1] \end{bmatrix} \in \mathbb{R}^{p+m\tau}, \tag{3}$$

not just $y[t]$. The state equation of (1) corresponding to the choices above is

$$\begin{cases} x_\tau[t+1] = A_\tau x_\tau[t] + B_\tau u[t] \\ y_\tau[t] = C_\tau x_\tau[t] \end{cases} \tag{4}$$

where

$$\left[\begin{array}{c:c} A_\tau & B_\tau \end{array}\right] := \left[\begin{array}{ccccc:c} A & B & 0 & \cdots & 0 & 0 \\ 0 & 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I & 0 \\ 0 & 0 & 0 & \cdots & 0 & I \end{array}\right], \quad C_\tau := \begin{bmatrix} C & D & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I \end{bmatrix}$$

Realization (4) inherits structural properties of the delay-free version of (1), meaning that the augmentation procedure has no redundancy.

Proposition 1. $\lambda \in \mathbb{C}$ is an unreachable mode of (A_τ, B_τ) iff it is an unreachable mode of (A, B) and is an unobservable mode of (C_τ, A_τ) iff it is an unobservable mode of (C, A).

Thus, $\boldsymbol{\mathcal{A}_1}$ implies that we can design a stabilizing controller for (4) under every τ. We follow the observer-based design procedure for that. Although (4) is a standard system, designing state-feedback and observer gains without accounting for the structure of its parameters would be a waste. The dimension of (4) might be very high if the delay is large, which would lead to numerical issues. So the trick is to exploit the structure of the parameters of (4) to end up with calculations independent of τ.

2.1 State feedback with partial pole placement

We start with discussing the design of the state-feedback gain $K_\tau \in \mathbb{R}^{m\times(n+m\tau)}$ for (4). It is readily seen that

$$\operatorname{spec}(A_\tau) = \operatorname{spec}(A) \cup \{0\}^{m\tau}.$$

By Proposition 1, the $m\tau$ deadbeat eigenvalues are reachable via B_τ and can thus be freely moved by state feedback. But a key question is whether we really need to move them. Eigenvalues at the origin are damped and have the fastest possible convergence. As such, having them is normally desirable. Moving non-zero modes to the origin is regarded as expensive and non-robust, for it might require high feedback gains. Yet if such modes are already present in the open-loop system, keeping them is both justified by performance considerations and economical (e.g. they would be kept by the "expensive control" LQR (Anderson and Moore, 1989, Sec. 6.2)). Thus, it is justified to design K_τ that moves only elements of the spectrum of A in $\operatorname{spec}(A_\tau)$.

The lemma below presents a possible approach to designing the state-feedback gain with the required property.

Lemma 1. If there is a rank-n matrix $M_\tau \in \mathbb{R}^{n\times(n+m\tau)}$ such that $M_\tau A_\tau = AM_\tau$ and $M_\tau B_\tau = B$, then $K_\tau = KM_\tau$ renders

$$\operatorname{spec}(A_\tau + B_\tau K_\tau) = \operatorname{spec}(A + BK) \cup \{0\}^{m\tau}$$

for every $K \in \mathbb{R}^{m\times n}$.

The formulation above appears somewhat peculiar. In principle, partial pole assignment would only require the row space of M_τ to match the left eigenspace of the movable part of $\operatorname{spec}(A_\tau)$. Yet the structure of A_τ and B_τ facilitates an explicit construction of the required M_τ even in this restrictive form.

Lemma 2. The matrix

$$M_\tau = \begin{bmatrix} A^\tau & A^{\tau-1}B & \cdots & AB & B \end{bmatrix}$$

satisfies the conditions of Lemma 1 iff $\boldsymbol{\mathcal{A}_2}$ holds.

Thus, we need to choose a state-feedback gain K rendering $A + BK$ Schur, which is always possible by $\boldsymbol{\mathcal{A}_1}$, and then use

$$K_\tau = K \begin{bmatrix} A^\tau & A^{\tau-1}B & \cdots & AB & B \end{bmatrix}.$$

Taking into account the structure of the state vector in (2), this gain results in the control law

$$u[t] = K_\tau x_\tau[t] = K\Big(A^\tau x[t] + \sum_{i=1}^{\tau} A^{i-1}Bu[t-i]\Big). \tag{5}$$

The signal in the parentheses above is the prediction of $x[t+\tau]$. As such, (5) is the discrete counterpart of the predictive control law, like that in (Furukawa and Shimemura, 1983, Eqns. (25) and (26)).

2.2 Reduced-order observer

Consider now the observer design for the case when the whole x_τ is not measurable. Like in the state-feedback case, Proposition 1 implies that a full-order observer can always be constructed for (4). But the structure of the parameters of (4) can again be used to simplify the result. A key observation now is that the last $m\tau$ components of the measured signal in (3) coincide with those of the state vector in (2). They are the history of the control signal u generated by the controller. We may then expect that no measurement noise corrupts these component of

the state in many applications. Hence, the use of a reduced-order observer is a natural choice for (4).

The specific form of reduced-order observer that we use is not the one with a minimal order, because we do not reduce measured parts of x from it. But this does not affect its design qualitatively. Repeating the derivation in (Friedland, 1986, Sec. 7.5) for this choice, we end up with the observer of x in the form

$$\hat{x}[t+1] = A\hat{x}[t] + Bu[t-\tau] - L(y[t] - C\hat{x}[t] - Du[t-\tau]) \quad (6)$$

whose $\hat{x}$ converges asymptotically to x, provided $L \in \mathbb{R}^{n\times p}$ is such that $A + LC$ is Schur (exists by $\boldsymbol{\mathcal{A}_1}$).

2.3 Observer-based controller

Finally, combining (5) with (6) we end up with the observer-based control law for (1) in the form

$$\begin{cases} \hat{x}[t+1] = (A+LC)\hat{x}[t] + (B+LD)u[t-\tau] - Ly[t] \\ u[t] = K\Big(A^{\tau}\hat{x}[t] + \sum_{i=1}^{\tau} A^{i-1}Bu[t-i]\Big) \end{cases} \quad (7)$$

which is the discrete counterpart of the observer-predictor controller in (Furukawa and Shimemura, 1983, Eqns. (36)–(38)). The control law (7) is stabilizing, with the closed-loop spectrum

$$\operatorname{spec}(A+BK) \cup \operatorname{spec}(A+LC) \cup \{0\}^{m\tau},$$

which follows by separation arguments as in (Friedland, 1986, Sec. 8.3).

Remark 1. (multiple input delays). The arguments above can be extended to processes of the form

$$x[t+1] = Ax[t] + \sum_{i=0}^{\tau} B_i u[t-i].$$

In the state-feedback part, we essentially only need to replace the matrix M_τ in Lemma 2 with

$$M_\tau = \begin{bmatrix} A^{\tau} & \bar{M}_\tau & \cdots & \bar{M}_2 & \bar{M}_1 \end{bmatrix},$$

where $\bar{M}_i = A^{-1}\bar{M}_{i+1} + A^{\tau-1}B_i$ for $i \in \mathbb{Z}_{1..\tau}$ with $\bar{M}_{\tau+1} = 0$, and design K on the basis of the pair $(A, \sum_{i=0}^{\tau} A^{\tau-i}B_i)$, rather than (A, B). Assumptions $\boldsymbol{\mathcal{A}_{1,2}}$ should then be adapted to these changes, of course. The observer part does not alter. ▽

3. NOISY MEASUREMENTS OF THE CONTROL INPUT

Although the assumption that u is measured perfectly can be justified in many situations, we may think of applications where this is not the case. For example, the control signal generated by the controller may be transmitted to the plant by a noisy communication channel. A way to formalize this is to assume that the control signal applied to the plant is

$$u[t] = u_{\mathrm{c}}[t] + v_u[t],$$

where u_{c} is the (measured) output of the controller and v_u is some noise. In such situations delayed u's in measured variable (3) become delayed $(u - v_u)$'s. And this alteration questions the use of observer (6) in the studied context, because reduced-order observers tend to be sensitive to measurement noise in reduced channels, see (Anderson and Moore, 1989, Sec. 7.2).

Our goal in this section is to shed light on this issue. To this end, we consider the system

$$\begin{cases} x[t+1] = Ax[t] + Bu[t-\tau] + B_w w[t] \\ y_x[t] = Cx[t] + Du[t-\tau] + D_w w[t] \\ y_u[t] = u[t] - D_v v[t] \end{cases} \quad (8)$$

and study the reconstruction of its state x_τ, as in (2), from measurements y_x and y_u. The signal w in the first two equations of (8) can be thought of as comprising plant disturbances and a measurement noise in the y-channel. Its introduction is required to render the state reconstruction problem well posed. The signal v, assumed to be independent of w, represents measurement imperfections in the u-channel and is the main focus below. It is convenient to use normalized w and v, so matrices B_w, D_w, and D_v reflect intensities of physical signals and their mutual relations (for parts of w), as well as design considerations. In particular, $v_u = D_v v$ for a fictitious unit-intensity v. Note that we consider only the measurement of $u[t]$ rather than of the whole τ-history of it as in (3). The reason is that we do not impose any structure on a reconstructor in the analysis below. In this case an explicit account for delayed versions of y_u would introduce redundancy, without affecting the performance.

By the reconstruction of x_τ we understand the generation of its estimate $\hat{x}_\tau$ from the measurements (y_x, y_u) in (8) by a linear system (filter) F. The goal is to choose F so that the reconstruction error $x_\tau - \hat{x}_\tau$ is small, in whatever sense. The size of the error can be quantified in terms of the *error system* G_{e}, which connects exogenous inputs, u, w, and v, with the error. To construct G_{e}, introduce the transfer functions

$$\begin{bmatrix} G_{xw} & G_{xu} \\ G_{yw} & G_{yu} \end{bmatrix} := \left[\begin{array}{c|cc} A & B_w & B \\ \hline I & 0 & 0 \\ C & D_w & D \end{array}\right], \quad W_\tau := \begin{bmatrix} z^{-\tau}I \\ \vdots \\ z^{-1}I \end{bmatrix}. \quad (9)$$

It is then readily verified that

$$x_\tau = \begin{bmatrix} G_{xu}z^{-\tau} & G_{xw} & 0 \\ W_\tau & 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ w \\ v \end{bmatrix}$$

and

$$\begin{bmatrix} y_x \\ y_u \end{bmatrix} = \begin{bmatrix} G_{yu}z^{-\tau} & G_{yw} & 0 \\ I & 0 & -D_v \end{bmatrix} \begin{bmatrix} u \\ w \\ v \end{bmatrix}.$$

Hence, the error system

$$G_{\mathrm{e}} = \begin{bmatrix} G_{xu}z^{-\tau} & G_{xw} & 0 \\ W_\tau & 0 & 0 \end{bmatrix} - F \begin{bmatrix} G_{yu}z^{-\tau} & G_{yw} & 0 \\ I & 0 & -D_v \end{bmatrix}.$$

We consider the following requirements in the design of F:

(1) both G_{e} and F itself are stable,
(2) F is strictly causal,
(3) the error $x_\tau - \hat{x}_\tau$ is decoupled from u,
(4) the H_2-norm $\|G_{\mathrm{e}}\|_2$ is minimized.

The first requirement should be evident, an unstable F cannot be safely implemented in open loop and the norm of an unstable error system is not well defined. The second and third requirements are motivated by the compatibility with the observer-based design in Section 2. The observer there is strictly causal, by construction, and the control signal does not affect the observation error. Moreover, the part of F acting on y_u must be strictly causal because the control signal $u[t]$ is only generated at the time instance t. An additional motivation for the third

requirement is that u has normally a different nature than disturbances, so mixing them in the analysis of the reconstruction performance is not justified. The fourth item is the standard steady-state Kalman filtering objective. If w and v are unit-variance white processes, then $\|G_e\|_2^2$ equals the steady-state variance of the reconstruction error.

To solve this problem, we need to assume that

$\boldsymbol{\mathcal{A}_3}$: $\begin{bmatrix} A-zI & B_w & BD_v \\ C & D_w & DD_v \end{bmatrix}$ has full row rank for all $|z|=1$.

This condition can always be warranted by a choice of B_w and D_w. Together with the detectability of (C,A), already included into $\boldsymbol{\mathcal{A}_1}$, assumption $\boldsymbol{\mathcal{A}_3}$ guarantees (Saberi et al., 2007, Thm. 4.79) that the discrete algebraic Riccati equation (DARE)

$$\begin{aligned} Y = {} & AYA' + B_w B_w' + BD_v D_v' B' \\ & - (B_w D_w' + BD_v D_v' D' + AYC')R^{-1} \\ & \times (D_w B_w' + DD_v D_v' B' + CYA'), \end{aligned} \quad (10)$$

where $R := D_w D_w' + DD_v D_v' D' + CYC'$, admits a stabilizing solution $Y = Y' \geq 0$ such that $R > 0$ and $A + LC$ is Schur, where $L := -(B_w D_w' + BD_v D_v' D' + AYC')R^{-1}$.

The main res

Theorem 1. If $\boldsymbol{\mathcal{A}_{1,3}}$ hold, then the strictly causal filter solving the problem of H_2-optimal reconstruction of the state x_τ of (8) is

$$\begin{cases} \hat{x}[t+1] = A\hat{x}[t] + By_u[t-\tau] - L\hat{y}[t] \\ \hat{u}[t-i] = y_u[t-i], \qquad \forall i \in \mathbb{Z}_{1..\tau} \end{cases} \quad (11)$$

where $\hat{y}[t] := y_x[t] - C\hat{x}[t] - Dy_u[t-\tau]$.

A remarkable outcome of Theorem 1 is that the optimal reconstruction of the τ-history of the control input from noisy measurements of u is still provided by the reduced-order observer having the same structure as that in §2.2. This is not an expectable result. Kalman filters may have a reduced-order observer form in some situations. However, this normally happens when noise vanishes in certain measurement channels, see (Friedland, 1986, Sec. 11.6). Our case is unusual, for the reduced-order structure is maintained for any noise intensity weight D_v. The latter still affects the optimal solution, via influencing the DARE (10) and thus the observer gain L.

Remark 2. (causal F_x). Allowing the first component of F, that generating x, to have a nonzero feedthrough part may change the solution structure. It can be shown that in this case the optimal filter is

$$\begin{cases} x_F[t+1] = Ax_F[t] + By_u[t-\tau] - L\hat{y}[t] \\ \hat{x}[t] = x_F[t] + YC'R^{-1}\hat{y}[t] \\ \hat{u}[t-i] = y_u[t-i], \qquad \forall i \in \mathbb{Z}_{1..\tau-1} \\ \hat{u}[t-\tau] = y_u[t-\tau] + D_v D_v' D' R^{-1}\hat{y}[t] \end{cases}$$

and the optimal cost attained by it is reduced by the quantity $\mathrm{tr}(R^{-1}(CY^2C' + D(D_vD_v')^2D'))$. If $D = 0$ we still have the reduced-order observer structure of the optimal strictly causal filter in (11), bar a change in generating $\hat{x}$. But if $D \neq 0$, the filter is no longer the reduced-order observer for the oldest estimated element of u, which is $u[t-\tau]$. ▽

4. CONCLUDING REMARKS

This note has studied discrete-time dead-time compensation from the perspective of the orthodox observer-based controller design. It has been shown that the discrete equivalent of the observer-predictor architecture can be derived from scratch via such arguments under mild assumptions. Specifically, only two choices had to be made. First, the deadbeat modes of the plant that originate in the delay element are not shifted by the state feedback. Second, delayed control inputs, which are parts of the state and measured directly, are excluded from the state observer (hence, the use of a reduced-order observer). The sensitivity of the last choice to measurement noise has also been investigated via posing and solving an H_2 filtering problem for the studied delayed setup. Remarkably, the H_2-optimal strictly causal filter is always of the reduced-order observer form, regardless the measurement noise intensity. This offers a solid justification for the observer-predictor setup, showing that it is not sensitive to measurement noise.

REFERENCES

Anderson, B.D.O. and Moore, J.B. (1989). *Optimal Control: Linear Quadratic Methods*. Prentice-Hall, Englewood Cliffs, NJ.

Artstein, Z. (1982). Linear systems with delayed control: A reduction. *IEEE Trans. Automat. Control*, 27(4), 869–879.

Fiagbedzi, Y.A. and Pearson, A.E. (1990). Output feedback stabilization of delay systems via generalization of the transformation method. *Int. J. Control*, 41(4), 801–822.

Friedland, B. (1986). *Control System Design: An Introduction to State-Space Methods*. McGraw-Hill, New York, NY.

Furukawa, T. and Shimemura, E. (1983). Predictive control for systems with time delay. *Int. J. Control*, 37(2), 399–412.

Krstic, M. (2009). *Delay Compensation for Nonlinear, Adaptive, and PDE Systems*. Birkhäuser, Boston, MA.

Manitius, A.Z. and Olbrot, A.W. (1979). Finite spectrum assignment problem for systems with delay. *IEEE Trans. Automat. Control*, 24, 541–553.

Meinsma, G. and Mirkin, L. (2005). H^∞ control of systems with multiple I/O delays via decomposition to adobe problems. *IEEE Trans. Automat. Control*, 50(2), 199–211.

Meinsma, G. and Zwart, H. (2000). On $\mathcal{H}_\infty$ control for dead-time systems. *IEEE Trans. Automat. Control*, 45(2), 272–285.

Mirkin, L. (2003). On the extraction of dead-time controllers and estimators from delay-free parametrizations. *IEEE Trans. Automat. Control*, 48(4), 543–553.

Mirkin, L. (2006). On the dead-time compensation from L^1 perspectives. *IEEE Trans. Automat. Control*, 51(6), 1069–1073.

Mirkin, L., Palmor, Z.J., and Shneiderman, D. (2011). Dead-time compensation for systems with multiple I/O delays: A loop shifting approach. *IEEE Trans. Automat. Control*, 56(11), 2542–2554.

Mirkin, L. and Raskin, N. (2003). Every stabilizing dead-time controller has an observer-predictor-based structure. *Automatica*, 39(10), 1747–1754.

Mirkin, L. and Zanutto, D. (2022). Dead-time compensation as an observer-based design. *IEEE Control Syst Lett*, 6, 1604–1609.

Saberi, A., Stoorvogel, A.A., and Sannuti, P. (2007). *Filtering Theory with Applications to Fault Detection, Isolation, and Estimation*. Birkhäuser, Boston, MA.

Smith, O.J.M. (1957). Closer control of loops with dead time. *Chem. Eng. Progress*, 53(5), 217–219.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Learning Hamiltonian Systems and Symmetries ⋆

Eva Dierkes * **Christian Offen** ** **Sina Ober-Blöbaum** **
Kathrin Flaßkamp ***

* *Center for Industrial Mathematics, University of Bremen, Germany, (e-mail: eva.dierkes@uni-bremen.de)*
** *Department of Mathematics, Paderborn University, Germany*
*** *Systems Modeling and Simulation, Saarland University, Germany*

Abstract: During the last years, Hamiltonian neural networks (HNN) have been introduced to incorporate prior physical knowledge when learning dynamical systems. Hereby, the symplectic system structure is preserved despite the data-driven modeling approach. However, preserving symmetries requires additional attention. In this research, we enhance the HNN with a Lie algebra framework to detect and embed symmetries in the neural network. This approach allows to simultaneously learn the symmetry group action and the total energy of the system.

Keywords: Physics informed neural networks, Hamiltonian systems, Hamiltonian learning, system identification, symmetry action identification

1. INTRODUCTION

Modeling mechanical system dynamics has a long history, many physical principles have been prescribed in detail and from different perspectives, e.g. the Hamiltonian and the Lagrangian viewpoint as well as the Newton-Euler modeling approach. However, also data-based modeling techniques have recently gained attention within this area. Reasons might range from bypassing complex physical modeling which requires domain-specific knowledge, e.g. in novel materials, to finding reduced-order models of e.g. multi-physics systems.

Mechanical systems are well-known to possess characteristic properties, such as symplecticity of the Hamiltonian/Lagrangian flow, symmetries which lead to the preservation of momentum maps, and energy-preservation in the absence of external forcing. In physics-based modeling, differential geometry allows to generate structure-preserving models, e.g. variational integrators provide discrete-time analogs to the original system structures. In data-based modeling, standard methods such as neural network approximations of vector fields or flows do not preserve these structures. Physics-informed neural networks were introduced by Raissi et al. (2019), in which the unknown parts of a PDE are learned by adding the expression of the PDE to the loss function and using automatic differentiation properties. The SINDy approach, proposed by Brunton et al. (2016) applies sparse regression to get a symbolic representation of the system's ODE based on suitable basis functions, without focusing on mechanical systems. However, Udrescu and Tegmark (2020) presented an approach where deep learning is used to find symmetries in the data in order to reduce the exponentially large search space of all possible basis functions.

⋆ E. Dierkes acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG) Project number 281474342.

The learning of Hamiltonian systems is addressed in Zhong et al. (2019). Within this work, the scalar valued Hamiltonian is learned by using neural networks to model the typical components of the Hamiltonian separately, such as the potential energy or the mass matrix. A more general approach was proposed by Greydanus et al. (2019) and named *Hamiltonian Neural Networks (HNN)*, since the Hamiltonian is learned such that the system's symplecticity and energy conservation are preserved by design. Extending this approach, Dierkes and Flaßkamp (2021) shows how to learn a symmetry-preserving Hamiltonian, if the symmetry is known a priori.

Rather than learning the continuous Hamiltonian, Offen and Ober-Blöbaum (2022) learn a Hamiltonian tailored to symplectic integration schemes. In this way, discretization errors in the integration step are eliminated and trajectory observations instead of observations of the Hamiltonian vector field can be used in the learning process. A similar strategy can be employed on the Lagrangian side (see Ober-Blöbaum and Offen, 2022).

Another approach to preserve symplectic structure when learning dynamical systems is to learn the system's Hamiltonian flow map by learning its generating function (cf. Rath et al., 2021) or by using symplectic neural networks (cf. Jin et al., 2020), where symplecticity is guaranteed by the network architecture. Moreover, symmetries have been embedded into learning vector fields using group integration matrix kernels (GIM kernels) and Gaussian Processes (cf. Ridderbusch et al., 2021). Lie algebra convolutional neural networks (cf. Dehmamy et al., 2021) can automatically discover symmetries and preserve them by using suitable localizations of the kernels present in CNNs and making them group invariant.

In this article, we will learn the Hamiltonian of a system and use a Lie algebra framework to detect and embed

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

symmetries into a neural network that models the Hamiltonian of a dynamical system. For proposed Lie group actions, such as the action of affine linear transformations, our network automatically detects subgroups under which the Hamiltonian function is invariant. This is achieved by learning the Hamiltonian along with a spanning set of generators of invariant vector fields and testing via a loss function whether the derivatives of H along the invariant vector fields vanish. Utilising Noether's theorem, the two ingredients, symmetries and symplectic structure, then allow us to identify integrals of motions. We exemplify the concept by the cart–pendulum example, for which we train a symmetry-symplecticity-preserving neural network.

2. LEARNING DYNAMICS AND SYMMETRIES

Consider an orientable, Riemannian manifold Q with phase space $(\mathcal{M}, \omega) = (T^*Q, \omega)$, where T^*Q denotes the cotangent bundle and ω the canonical symplectic structure. Let $\mathfrak{X}(\mathcal{M})$ denote the set of vector fields on $\mathcal{M}$.

In the following, we develop a loss function to learn a Hamiltonian $H\colon \mathcal{M} \to \mathbb{R}$ modelled as a neural network based on observations $(\dot{z}^{(k)})_{k=1}^N$ of a Hamiltonian vector field $X_H \in \mathfrak{X}(\mathcal{M})$ at positions $(z^{(k)})_{k=1}^N \subset \mathcal{M}$. If $Q = \mathbb{R}^n$, $X_H(q,p) = (\nabla_p H(q,p), -\nabla_q H(q,p))^\top$ in standard (Darboux) coordinates $z = (q,p) \in \mathcal{M} \cong \mathbb{R}^{2n}$.

Our loss function will guide the network towards a symmetric Hamiltonian function: given a symplectic action of a Lie group G on $\mathcal{M}$, the minimization procedure identifies a subgroup of G which acts by symmetries. Once the symmetries are known, conserved quantities can be derived by Noether's theorem. If, for instance, the Hamiltonian is invariant under translations, i.e. the directional derivative $\nabla_{(w,0)} H(q,p) = \sum_{j=1}^n w^j \frac{\partial H}{\partial q^j}(q,p)$ vanishes for a translation direction $w \in Q = \mathbb{R}^n$ and all $(q,p) \in \mathcal{M} \cong \mathbb{R}^{2n}$, then the quantity $I(q,p) = w \cdot p$ is conserved under motions.

The loss function ℓ consists of a dynamical part ℓ_{dynamics} and a part related to symmetries ℓ_{sym}. We have

$$\ell_{\text{dynamics}} = \sum_{k=1}^N \|\dot{z}^{(k)} - X_H(z^{(k)})\|^2_{T\mathcal{M}},$$

which corresponds to HNN. In the following, we show how to construct ℓ_{sym} to a given Lie group, whose actions are possible symmetries.

2.1 Background on Lie-group actions

Let us briefly introduce Lie group actions and invariant vector fields. For details we refer to the book by Marsden and Ratiu (1999).

For a Lie group G, let $\mathfrak{g}$ denote its Lie algebra and $\exp\colon \mathfrak{g} \to G$ the exponential map. Consider a symplectic group action $L\colon G \to \mathrm{Symp}(\mathcal{M})$, $g \mapsto L_g$. Here $\mathrm{Symp}(\mathcal{M})$ denotes the group of symplectic diffeomorphisms. If $Q = \mathbb{R}^n$, as before, this means that L_g is required to preserve the symplectic structure, i.e. $DL_g^T J DL_g = J$, where DL_g is the Jacobian matrix of L_g and J is the symplectic structure matrix $J = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}$ with the n-dimensional identity matrix I_n. For $v \in \mathfrak{g}$ the left invariant vector field $\hat{v} \in \mathfrak{X}(\mathcal{M})$ is defined by

$$\hat{v}_z = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} L_{\exp(tv)}(z) \in T_z\mathcal{M}, \quad z \in \mathcal{M}.$$

These vector fields can be thought of as infinitesimal actions of the Lie group G on $\mathcal{M}$. Invariance of H can be tested by computing directional derivatives of H in the directions given by these vector fields.

Example 1. (Affine linear transformations). Let $Q = \mathbb{R}^n$ and $\mathcal{M} = T^*Q \cong \mathbb{R}^{2n}$ with Darboux coordinates (q,p). The group of affine transformations $G = \mathrm{Aff}(Q)$ on Q can be represented as

$$G = \left\{ \begin{pmatrix} A & w \\ 0 & 1 \end{pmatrix} \,\middle|\, A \in \mathrm{Gl}(\mathbb{R}, n),\, w \in \mathbb{R}^n \right\},$$

where the group operation is matrix multiplication and $\mathrm{Gl}(\mathbb{R}, n)$ denotes the general linear group. Its Lie algebra is

$$\mathfrak{g} = \left\{ \begin{pmatrix} M & w \\ 0 & 0 \end{pmatrix} \,\middle|\, M \in \mathrm{Mat}(\mathbb{R}, n),\, w \in \mathbb{R}^n \right\}$$

and $\exp\colon \mathfrak{g} \to G$ is the matrix exponential. G acts on $\mathcal{M} = T^*Q$ as

$$L_{(A,w)}(q,p) = \left(A^{-1}(q-w),\, A^\top p\right),$$

where $g = (A,w)$ is a shorthand for $\begin{pmatrix} A & w \\ 0 & 1 \end{pmatrix}$. The action is symplectic as it arises as the cotangent lifted action of the action $G \to \mathrm{Diff}(Q)$, $(A,w) \mapsto L^Q_{(A,w)}$, with $L^Q_{(A,w)}(q) = Aq + w$. The invariant vector field $\widehat{v} = \widehat{(M,w)}$ corresponding to the action $L\colon G \to \mathrm{Symp}(\mathcal{M})$ at $(q,p) \in \mathcal{M}$ is given as

$$\widehat{v}_{(q,p)} = \widehat{(M,w)}_{(q,p)} = \left(-Mq - w,\, M^\top p\right).$$

Vector fields can be interpreted as derivations: if $\widehat{v} = \widehat{(M,w)}_{(q,p)}$ is applied to $H\colon \mathcal{M} \to \mathbb{R}$ we obtain the directional derivative

$$\begin{aligned}\widehat{v}_{(q,p)}(H) &= \widehat{(M,w)}_{(q,p)}(H) = \nabla_{\widehat{(M,w)}} H(q,p) \\ &= (-Mq - w)^\top \nabla_q H(q,p) + (M^\top p)^\top \nabla_p H(q,p).\end{aligned}$$

Example 2. (Translations). Restricting to translations $G \cong (\mathbb{R}^n, +)$, we have $\widehat{v}_{q,p} = \widehat{w}_{q,p} = (-w, 0)$ with $\widehat{w}_{q,p}(H) = -w^\top \nabla_q H(q,p)$.

2.2 Incorporation of symmetry into the loss function

Simultaneously to learning H, we aim to learn an orthonormal basis $v^{(1)}, \ldots, v^{(K)}$ of $\mathfrak{g}$ spanning a subspace $V \subset \mathfrak{g}$ such that H is invariant under actions with elements g of the subgroup $\exp(V) \subset G$.

Let $\mathcal{M}^{\mathrm{o}} \subset \mathcal{M}$ be open, pre-compact, and contain the parts of interest of the phase space $\mathcal{M}$. Let dvol be a volume form on $\mathcal{M}$. For $k = 1, \ldots, K$ define

$$\ell_{\text{sym}}^{(k)} = \frac{1}{\mathrm{dvol}(\mathcal{M}^{\mathrm{o}})} \int_{\mathcal{M}^{\mathrm{o}}} |\hat{v}^{(k)}(H)|^2 \,\mathrm{dvol}. \tag{1}$$

Here $\hat{v}^{(k)}$ denotes the invariant vector field to v. The term $\ell_{\text{sym}}^{(k)}$ measures how invariant H is under actions with group elements of $\exp(tv^{(k)} | t \in \mathbb{R})$. Equip $\mathfrak{g}$ with an inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Given weights $\alpha^{(k)}, \beta^{(k)} > 0$ define

$$\ell_{\text{sym}} = \sum_{k=1}^{K} \left(\ell_{\text{sym}}^{(k)} + \alpha^{(k)} |\|v^{(k)}\| - 1|^2 + \beta^{(k)} \sum_{s=1}^{k-1} \langle v^{(k)}, v^{(s)} \rangle \right). \tag{2}$$

The last two terms of ℓ_{sym} measure the orthonormality of the spanning set $v^{(1)}, \dots, v^{(K)}$ while the first term measures how well infinitesimal actions by elements of $\exp(V)$ preserve H.

Example 3. If we look for a 1-dimensional subgroup of $G = (\mathbb{R}^n, +)$, where G acts by translations on $Q = \mathbb{R}^n$ (example 2), we have for a Lie algebra element $v = w \in \mathfrak{g}$

$$\ell_{\text{sym}} = \frac{1}{\operatorname{dvol}(\mathcal{M}^{\circ})} \int_{\mathcal{M}^{\circ}} \left| \sum_{j=1}^{n} w^j \frac{\partial H}{\partial q^j}(q,p) \right|^2 \mathrm{d}q\, \mathrm{d}p,$$

where $\mathrm{d}q\, \mathrm{d}p = \mathrm{d}q^1 \dots \mathrm{d}q^n\, \mathrm{d}p_1 \dots \mathrm{d}p_n$.

2.3 Training and further remarks

The Hamiltonian H and the spanning set $v^{(1)}, \dots, v^{(K)}$ can now be learned using the loss function $\ell = \ell_{\text{dynamics}} + \ell_{\text{sym}}$, where H is modelled as a neural network and $v^{(1)}, \dots, v^{(K)}$ are additional parameters. Alternatively to learning all at once, the training can be performed with a low value for K first. (For $K = 0$ our method corresponds to HNN.) Then K is increased and the training repeated using the pre-trained network and $v^{(1)}, \dots, v^{(K-1)}$ as priors.

Remark 4. The integral in (1) can be approximated by averaging the integrand over a few points in the phase space $\mathcal{M}$, randomly drawn in each epoch of the minimization procedure.

Remark 5. Let $\mu\colon \mathcal{M} \to \mathfrak{g}^*$ be the momentum map of the considered group action and $\langle \cdot, \cdot \rangle_{\text{pair}}$ the dual pairing of the Lie algebra $\mathfrak{g}$ and its dual $\mathfrak{g}^*$. Provided that H accurately describes the Hamiltonian of the system and $\exp(V)$ its symmetry group, then the components $\mu^{(j)} = \langle \mu, v^{(j)} \rangle_{\text{pair}}$ of the momentum map are conserved quantities of the system's Hamiltonian motions. In case of example 2, the conserved quantities are $(q,p) \mapsto w_1^{(j)} p_1 + \dots + w_n^{(j)} p_n$, where $v^{(j)} = (w_1^{(j)}, \dots, w_n^{(j)})$.

Remark 6. Our framework works without any changes for general symplectic group actions, i.e. for actions which are not cotangent lifted actions. Indeed, the Riemannian symplectic manifold $(\mathcal{M}, \omega)$ does not need to have the structure of a cotangent bundle.

3. NUMERICAL RESULTS

The framework introduced in the previous section is now applied to the example of a planar pendulum mounted on a cart (cf. e.g. Bloch, 2003). Since the dynamics are translational invariant, i.e. independent of the cart's position, we restrict ourselves to learn the correct translation group.

The generalized coordinates are $q = (s, \varphi)$, where s is the position of the cart and φ is the angle to the upper vertical of the pendulum. Since this system is well studied, we can use the true Hamiltonian for, firstly, generating data points and, secondly, evaluating the performance of our data-based approach. It is given by

$$H(\varphi, p_s, p_\varphi) = \frac{a p_s^2 + 2 b p_s p_\varphi \cos\varphi + c p_\varphi^2}{2ac - b^2 \cos^2\varphi} - D \cos\varphi, \tag{3}$$

with the constants $a = ml^2$, $b = ml$, $c = M + m$ and $D = -mgl$, using l as the length and m as the mass of the pendulum, M corresponds to the mass of the cart and $g = 9.81$ to gravitation. For simplicity, all remaining constants are set to one, i.e. $m = l = M = 1$.

To generate a data set, 1500 trajectories are generated using random initial values with $|s| < 5$, $|\varphi| < \pi$, $|p_s| < 1$ and $|p_\varphi| < \pi$. Each trajectory has a length of 3 s and a sampling rate of 15 Hz. A fourth-order Runge–Kutta scheme is applied with a low error tolerance of 1e−10 and using the (exact) vector field induced by H (cf. Runge, 1895). To each sample of the resulting data set, Gaussian noise with $\sigma^2 = 1\text{e}{-2}$ is added.

The network architecture is similar to the proposed network by Greydanus et al. (2019), which consists of 2 layers with 512 neurons each, using tanh as an activation function. For the training, an Adam optimizer (proposed by Kingma and Ba (2017)) and a reduce on plateau learning rate scheduler (cf. Ayyadevara and Reddy, 2020) with an initial learning rate of 1e−3, a reduction factor of 0.8 is used. For the training of our proposed symmetry HNN (SymHNN) we set $K = 2$ and in (2) the scaling factor $\beta^{(k)}$ is set to 0.001 for all k. This allows the network to choose $v^{(1)} = v^{(2)}$ in case that the translational invariant only exists in one direction.

For the numerical comparison a dense neural network (DenseNN), without learning the Hamiltonian but directly learning the dynamics, and a HNN are trained. For each model the same Runge–Kutta scheme, as for data generation, is used to evaluate the performance. The resulting trajectories for the states are shown in Figure 1 for (randomly chosen) initial configuration $[q,p] = [3.89, 3.68, -0.95, -1.67]$. The true Hamiltonian (3) is evaluated for the states of Figure 1 and shown in

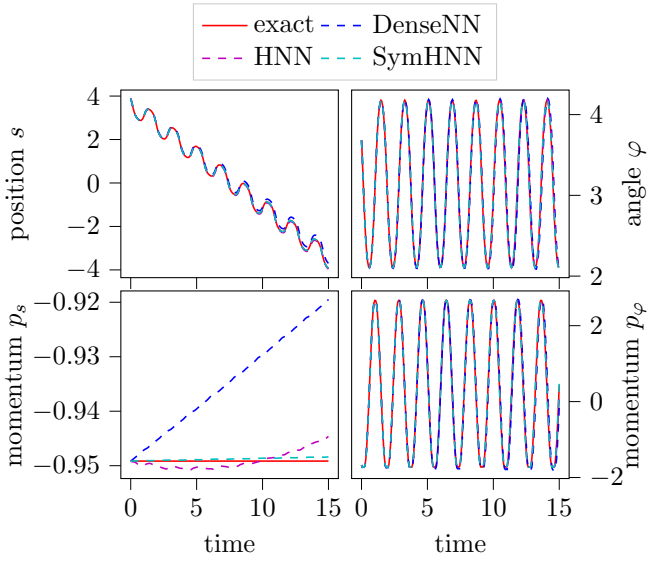


Fig. 1. Evaluated states of different (learned) models for the cart-pendulum.

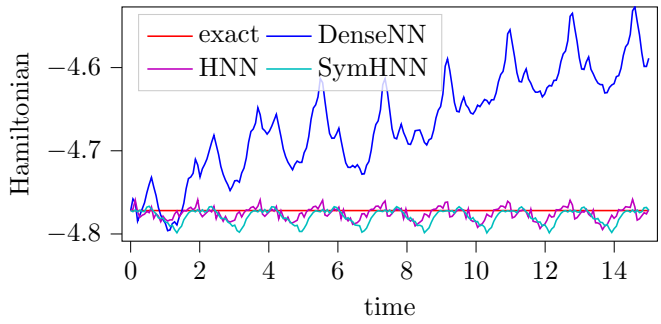


Fig. 2. True Hamiltonian evaluated for the states shown in Figure 1.

Figure 2. One can observe that the Hamiltonian is learned as successfully with SymHNN as with HNN.

Since this example is known to be invariant in the position s, the momentum p_s is conserved, which leads to a constant p_s for each trajectory. In Figure 1 it can be seen that the momentum for the DenseNN and the HNN are not conserved, whereas the momentum for our proposed SymHNN is preserved more effectively. To confirm this, the values for $\ell_{\text{sym}}^{(k)}$ are computed with the reference value $v = [1, 0]$ for all models. An improvement by a factor of 10 is observed (DenseNN: 0.0125, HNN: 0.0196 and SymHNN: 0.0012).

It should be highlighted that SymHNN learns correctly to choose $v^{(1)} = v^{(2)} = [0.990, -1.775\text{e}{-05}]$, since this example only has one translational invariant in s-direction. The learned v is close to the reference value of $[1, 0]$. However, since the integration in (1) is not performed exactly and the symmetry losses $\ell_{\text{sym}}^{(k)}$ are not precisely zero, the conjugate momentum p_s is not exactly but only approximately conserved in Figure 1.

4. CONCLUSION

We propose a neural network approach for simultaneously learning a system's Hamiltonian and its symmetries based on trajectory data. To preserve the symplectic structure encoded in the data, the NN is trained to learn the Hamiltonian, since the vector field can then be generated via automatic differentiation (cf. Greydanus et al., 2019; Dierkes and Flaßkamp, 2021). Extending this approach, which as coined the HNN method, we simultaneously identify inherent symmetries. Since we focus on the Hamiltonian setting, according to Noether's theorem, symmetries present themselves in term of invariances. However, we go another step forward and learn basis vectors of a Lie algebra subspace. These define the Lie subgroup, which belongs to actions (e.g. affine linear transformations) leaving the Hamiltonian invariant. In other words, we identify the integrals of motions such as the cart position in the studied cart-pendulum example. Future work will address the combination with symplectic discretization, e.g. via symplectic partitioned Runge-Kutta schemes. A discrete-time variant of this approach might directly learn modified Hamiltonian and discerete-time symmetries (cf. Offen and Ober-Blöbaum, 2022; Ober-Blöbaum and Offen, 2022). Symmetry in Hamiltonian systems give rise to relative equilibria, to which the learning framework can be extended in future.

REFERENCES

Ayyadevara, V. and Reddy, Y. (2020). *Modern Computer Vision with PyTorch: Explore deep learning concepts and implement over 50 real-world image applications*. Packt Publishing.

Bloch, A.M. (2003). *Nonholonomic mechanics and control*. Springer.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA*, 113(15), 3932–3937. doi: 10.1073/pnas.1517384113.

Dehmamy, N., Walters, R., Liu, Y., Wang, D., and Yu, R. (2021). Automatic Symmetry Discovery with Lie Algebra Convolutional Network. *arXiv preprint arXiv:2109.07103*.

Dierkes, E. and Flaßkamp, K. (2021). Learning Hamiltonian Systems considering System Symmetries in Neural Networks. *IFAC-PapersOnLine*, 54(19), 210–216.

Greydanus, S., Dzamba, M., and Yosinski, J. (2019). Hamiltonian Neural Networks. In *Advances in Neural Information Processing Systems*, 15353–15363. Canada.

Jin, P., Zhang, Z., Zhu, A., Tang, Y., and Karniadakis, G.E. (2020). Sympnets: Intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems. *Neural Networks*, 132, 166–179. doi: 10.1016/j.neunet.2020.08.017.

Kingma, D.P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Marsden, J.E. and Ratiu, T.S. (1999). *Introduction to mechanics and symmetry*, volume 17 of *Texts in Applied Mathematics*. Springer, 2nd edition.

Ober-Blöbaum, S. and Offen, C. (2022). Variational integration of learned dynamical systems. *arXiv preprint arXiv:2112.12619*.

Offen, C. and Ober-Blöbaum, S. (2022). Symplectic integration of learned Hamiltonian systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(1), 013122. doi:10.1063/5.0065913.

Raissi, M., Perdikaris, P., and Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378, 686–707. doi:10.1016/j.jcp.2018.10.045.

Rath, K., Albert, C.G., Bischl, B., and von Toussaint, U. (2021). Symplectic Gaussian process regression of maps in Hamiltonian systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(5), 053121. doi: 10.1063/5.0048129.

Ridderbusch, S., Offen, C., Ober-Blöbaum, S., and Goulart, P. (2021). Learning ODE models with qualitative structure using Gaussian processes. *2021 60th IEEE Conference on Decision and Control (CDC)*. doi: 10.1109/cdc45484.2021.9683426.

Runge, C. (1895). Über die numerische Auflösung von Differentialgleichungen. *Mathematische Annalen*, 46, 167–178. doi:10.1007/BF01446807.

Udrescu, S.M. and Tegmark, M. (2020). AI Feynman: a Physics-Inspired Method for Symbolic Regression. *Sci. Advances*. ArXiv: 1905.11481.

Zhong, Y.D., Dey, B., and Chakraborty, A. (2019). Symplectic ode-net: Learning Hamiltonian dynamics with control. *arXiv preprint arXiv:1909.12077*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Time minimal control of multi-level quantum systems

Rachida El Assoudi-Baikari * **I. Edouard Zibo** **

* *Normandie Université, INSA. Avenue de l'Université, 76801 Saint Etienne du Rouvray, France. (e-mail: rachida.el-assoudi@insa-rouen.fr).*
** *Normandie Université, INSA. Avenue de l'Université, 76801 Saint Etienne du Rouvray, France. (e-mail: edouard.zibo@insa-rouen.fr)*

Abstract: We study time minimal control problem for quantum systems whose dynamics are governed by the Bloch equation with interaction. The dynamics of the quantum systems are analyzed as affine control systems on the Bloch ball using parametrizations of the density matrix. The influence of Coulomb energies during a process of population transfer for a quantum system with several energy levels is shown and time minimal trajectories are given.

Keywords: Optimal control, Quantum dots, Bloch models, Elliptic functions.

1. INTRODUCTION

In the last few decades control of quantum systems has been widely studied from both theoretical and interdisciplinary points of view. Recently, there has been a growing interest in optimal control of quantum systems because of their applications to biology, physics, chemistry and quantum computing. Also, studies on the manipulation of quantum systems have given rise to several models allowing the description of certain physical phenomena. For example, the Landau-Zener and Lindblad equations which describe the evolution of quantum systems interacting with their environment are given in Morzhin et al. (2021). Other applications include control of spin dynamics by magnetic fields in nuclear magnetic resonance in Ernst et al. (1987). We are interested in the time minimal control problem for quantum systems at several energy levels given by quantum dots and whose dynamics are described by the Bloch equation with interaction, taking Coulomb parameters into account.
The control of quantum systems and time minimum population transfer problem has attracted interest of many authors. For example, in Sugny et al. (2007) the problem of achieving the optimal synthesis of a dissipative quantum system at two energy levels is treated. In Boscain et al. (2006) the authors consider the time minimum population transfer problem for a spin particle driven by a magnetic field on the Bloch sphere where the population dynamics are influenced by a parameter which depends on the maximum amplitude of the control field and energy levels. In Bonnard et al. (2009) and in Lapert et al. (2013) the problem of time minimal control with dissipative terms, and the damping effects which act on the dynamics of the populations is studied. In Khaneja et al. (2001) the authors deal with the problem of finding the time optimal control of spin quantum systems to produce unit operators.

In this paper we are interested in the study of time minimal control problem for quantum systems given by quantum dots, which are nonlinear Bloch models with interaction. We consider a Bloch model given by the complex matrix differential equation:

$$i\hbar\dot{\rho} = [H + V(\rho), \rho],$$

where state ρ is the density matrix, H is the Hamiltonian and $V(\rho)$ is the Coulomb interaction matrix depending on Coulomb parameters. A parameterization of density matrix ρ verifying physical properties, allows to write the Bloch model as an affine control system with some constraints. It is well-known that for a two-level system, the parameterization of ρ is in bijection with the Bloch vector, but for a system with three levels there exist different geometries and the parametrization is not unique (see Brning et al. (2012)). The influence of Coulomb parameters on time minimal trajectories for two-level systems is studied in Zibo et al. (2020).

Here we consider a four-level quantum system, to simplify the study of the system a choice of block parameterization of two-level systems is made. Then by a choice of the energies of the free Hamiltonian of the system, we show a symmetry in the dynamics of the system which evolves on a sphere of R^6. Also, we highlight the Coulomb interaction between two Bloch vectors. We apply Pontryagin's Maximum Principle to determine the minimum transfer time trajectories from an initial to a final pur or mixed states. We show that the trajectories of time minimal control problem are expressed as elliptic functions. Many authors (Brockett et al. (1993), D'Alessandro et al. (2001), Jurdjevic (2001) and Yuan et al. (2007)) were interested in elliptic functions and in more interesting cases, the solutions are expressible in terms of elliptic functions. In El Assoudi-Baikari et al. (2021) and (2016), it is shown that the solutions are not always elliptic, this depends of some geometrical parameters characterizing the state space.

2. MAIN RESULTS

We present briefly the nonlinear Bloch model. Parametrization of the density matrix give an affine control system that we will consider to study the transfer problem in minimal time using Pontryagin's Maximum Principle.

2.1 Nonlinear Bloch model

We present a nonlinear Bloch model that takes into account Coulomb effects in a quantum system called a quantum dot. In Bidgaray et al. (2014) a model of quantum dots which contain N^c energy levels in the conduction band and N^v energy levels in the valence band is provided. This modeling leads to a nonlinear Bloch type equation whose nonlinear terms come from the Coulomb interaction. This model is given by:

$$i\hbar\dot{\rho} = [H + V(\rho), \rho]$$

where $\hbar$ is constant of Planck, (we assume $\hbar = 1$), ρ is the density matrix, that is Hermitian, positive semi-definite and it has a trace equal to one. The state ρ represents pure and mixed quantum states if $trace(\rho^2) \leq 1$.
The matrix $H{=}H_0 + H_L$ is the total Hamiltonian, H_0 is the free energy Hamiltonian in the quantum dot and H_L is the laser-quantum dot interaction Hamiltonian. H_0 is a diagonal matrix whose coefficients are $(E_i^c)_{i\in I^c}$ and $(E_j^v)_{j\in I^v}$, real numbers representing the free energies of electrons in the conduction band and the valence band, respectively. The inter-band transition frequency of the quantum dots is the energy $E_{kj} = E_j^c - E_k^v$. Moreover, taking the Coulomb interaction into account modifies the energy levels, for example by causing them to tend downwards, and $E_{kj} = E_j^c - E_k^v - R_{jkjk}^{c-v}$, where R_{jkjk}^{c-v} account for the Coulomb interactions called Coulomb parameters (see Bidgaray et al. (2014)).
$H_L = \begin{pmatrix} 0 & E(t)M \\ \overline{E(t)M} & 0 \end{pmatrix}$, where $E(t)$ is the electric field and M the dipolar matrix.
The Coulomb interaction matrix $V(\rho)$ is Hermitian, depending on Coulomb parameters (see Bidgaray et al. (2014) and Zibo (2021)).

2.2 Parametrization of a model with four energy levels.

In this section we consider the evolution of a quantum dot with four energy levels described by the transitions between the level 1 in the conduction band and the level 2 in the valence band and the transitions between the level 2 in the conduction band and level 1 in the valence band. We consider a Hamiltonian matrix, density matrix and Coulomb interaction matrix per block of two matrices. Here we give a parametrization of this model as an affine control system with two controls.

We choose the Hamiltonian of the form $H = \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix}$, with $H_1 = \begin{pmatrix} E_1^c & \omega_1(t) \\ \omega_1(t) & E_2^v \end{pmatrix}$ and $H_2 = \begin{pmatrix} E_2^c & \omega_2(t) \\ \omega_2(t) & E_1^v \end{pmatrix}$, where E_1^c, E_2^v, E_2^c and E_1^v are the free energies for the conduction and valence bands, and $\omega_i(t)$, $i \in \{1, 2\}$ are real bounded valued controls.

$\rho = \begin{pmatrix} \rho_{11}^c & \rho_{12}^{cv} & 0 & 0 \\ \rho_{21}^{vc} & \rho_{22}^v & 0 & 0 \\ 0 & 0 & \rho_{22}^c & \rho_{21}^{cv} \\ 0 & 0 & \rho_{12}^{vc} & \rho_{11}^v \end{pmatrix}$ et $V(\rho) = \begin{pmatrix} V_{11}^c & V_{12}^{cv} & 0 & 0 \\ V_{21}^{vc} & V_{22}^v & 0 & 0 \\ 0 & 0 & V_{22}^c & V_{21}^{cv} \\ 0 & 0 & V_{12}^{vc} & V_{11}^v \end{pmatrix}$.
We write
$\rho = \begin{pmatrix} \rho(x) & 0 \\ 0 & \rho(y) \end{pmatrix}$, $\rho(x) = \frac{1}{4}\begin{pmatrix} 1+x_3 & x_1 - ix_2 \\ x_1 + ix_2 & 1 - x_3 \end{pmatrix}$.
where $x = (x_1, x_2, x_3)^t$, $y = (x_4, x_5, x_6)^t$ are Bloch vectors.

In Zibo (2021), we give in detail the associated Coulomb interaction matrix $V(\rho)$, we calculate its coefficients using the the properties of Coulomb parameters. We obtain:

$V_{11}^c = \frac{1}{2}(c_0 - c_1) - \frac{3}{4}(c_2 + c_3) - \frac{c_3}{4}x_3 + (\frac{1}{2}(c_0 - c_1) - \frac{c_2}{4})x_6$
$V_{22}^v = \frac{3}{2}(c_1' - c_0') + \frac{1}{4}(c_3 + c_4) + \frac{c_3}{4}x_3 + (\frac{1}{2}(c_1' - c_0') + \frac{c_4}{4})x_6$
$V_{12}^{cv} = -\frac{1}{4}(c_3x_1 + c_5x_4) + \frac{i}{4}(c_3x_2 + c_5x_5)$
$V_{22}^c = \frac{1}{2}(c_0 - c_1) - \frac{3}{4}(c_3 + c_4) + (\frac{1}{2}(c_0 - c_1) - \frac{c_4}{4})x_3 - \frac{c_3}{4}x_6$
$V_{11}^v = \frac{3}{2}(c_1' - c_0') + \frac{1}{4}(c_3 + c_2) + (\frac{1}{2}(c_1' - c_0') + \frac{c_2}{4})x_3 + \frac{c_3}{4}x_6$
$V_{21}^{cv} = -\frac{1}{4}(c_3x_4 + c_5x_1) + \frac{i}{4}(c_3x_5 + c_5x_2)$.
where c_j, c_j', for $0 \leq j \leq 5$ are constant depending on the Coulomb parameters.

Also, by identifying the matrices $\dot{\rho} = \dot{\rho}(x, y)$ and the bracket $-i[H + V(\rho), \rho]$ we obtain following system that highlights the interaction between the two blocks of ρ.

$\dot{x}_1 = a_1x_2 - cx_3x_5 + bx_2x_6$,
$\dot{x}_2 = -a_1x_1 + cx_3x_4 - bx_1x_6 - 2\omega_1x_3$,
$\dot{x}_3 = c(x_1x_5 - x_2x_4) + 2\omega_1x_2$,
$\dot{x}_4 = a_2x_5 - cx_2x_6 + bx_3x_5$,
$\dot{x}_5 = -a_2x_4 + cx_1x_6 - bx_3x_4 - 2\omega_2x_6$,
$\dot{x}_6 = c(x_2x_4 - x_1x_5) + 2\omega_2x_5$,
with reel constants a_1, a_2, b, c given by Coulomb parameters and the transition frequencies.

It is easy to verify that for any control functions ω_1, ω_2, we have $x_1\dot{x}_1 + x_2\dot{x}_2 + x_3\dot{x}_3 = 0$ and $x_4\dot{x}_4 + x_5\dot{x}_5 + x_6\dot{x}_6 = 0$ which implies $\sum_{j=1}^6 x_j\dot{x}_j = 0$. Hence for $x(0) = (0, 0, 1)^t$ and $y(0) = (0, 0, 1)^t$, one has that $\|x(t)\|^2 = \|x(0)\|^2 = 1$ and $\|y(t)\|^2 = \|y(0)\|^2 = 1$, and then $x(t)$, $y(t) \in S^2 \subset R^3$. Set $z = \begin{pmatrix} x \\ y \end{pmatrix}$, clearly $\|z(t)\|^2 = \|z(0)\|^2 = 2$. So $z(t)$ belongs to the sphere of radius $\sqrt{2}$ in R^6.
Notice that $x(t)$ and $y(t)$ do not independently evolve, there is an interaction between these two Bloch vectors.

2.3 Symmetry and interaction in the dynamics

We choose the free Hamiltonian $H_0 = diag(E_1^c, E_2^v, E_2^c, E_1^v)$ such that the two energy differences $E_1^c - E_2^v$ and $E_2^c - E_1^v$ being equal and then we get that $a_1 = a_2 = a$. In this case the following proposition shows a symmetry in the dynamics of the Bloch vectors $x(t)$ and $y(t)$ and that the linear part of the above system corresponds to a two-level system for each Bloch vector $x(t)$ and $y(t)$ and the nonlinear part representing the Coulomb interaction, is a vector product of the two vectors $x(t)$ and $y(t)$. Set

$A = \begin{pmatrix} 0 & a & 0 \\ -a & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$ and $C = \begin{pmatrix} c & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & b \end{pmatrix}$.
Set $u_1(t) = 2\omega_1(t)$ and $u_2(t) = 2\omega_2(t)$ and assume $|u_i| \leq 1$.

Proposition 1 The dynamic of this quantum system with four energy levels, when $a_1 = a_2 = a$, is given by

$$(\Sigma) \quad \dot{z} = \begin{pmatrix} A + u_1 B & 0 \\ 0 & A + u_2 B \end{pmatrix} z + \begin{pmatrix} x \wedge Cy \\ y \wedge Cx \end{pmatrix}.$$

2.4 Time minimal trajectories

We present some basic notions of the optimal control problem (see for more details Pontryagin et al. (1962) and Boscain et al. (2004)). Consider an affine control system with m control functions $\dot{z} = F(z) + \sum_{i=1}^{m} u_i G_i(z)$, where $z \in R^n$ and $F, G_1, ..., G_m$ are smooth vector fields and control functions $u = (u_1, ..., u_m)^t$ are bounded measurable functions, with values in a domain $\mathcal{U} \subset R^m$. An optimal control problem consists to find a trajectory $z(\cdot)$ associated with a control $u(\cdot)$, solution of this system verifying $z(0) = z_0$ and $z(T) = z_1$ and minimizing a functional cost $C(u) = \int_0^T L(z(t), u(t))dt$, where $L : R^n \times R^m \longrightarrow R$ is a criterion function. The pair $(z(\cdot), u(\cdot))$ is called optimal. If $L(z(t), u(t)) = 1$, then $C(u) = T$ and we have a time minimal control problem.

We apply Pontryagin's Maximum Principle (PMP) which is a generalization of the Hamiltonian formulation of the classical calculus of variations and it gives necessary conditions that a trajectory must satisfy in order to minimize a functional cost. The Hamiltonian function associated to the system with the cost $C(u) = T$ is defined as follows for all $(z, p, u) \in T^*R^n \times \mathcal{U}$, where $p(t)$ is a row vector (the adjoint vector), and p_0 a constant, $p_0 \leq 0$.

$$\mathcal{H}(z, p, u) = p(t)(F(z(t)) + \sum_{i=1}^{m} u_i G_i(z(t))) + p_0$$

Pontryagin's Maximum Principle says that if a pair $(z(\cdot), u(\cdot))$ is optimal, then there exists a non zero absolutely continuous function $p(\cdot) : [0, T] \longrightarrow T^*_{z(t)}R^n$, and a constant $p_0 \leq 0$, such that for almost all $t \in [0, T]$ we have

$\dot{z}(t) = \frac{\partial \mathcal{H}}{\partial p}(z(t), p(t), u(t))$, $\dot{p}(t) = -\frac{\partial \mathcal{H}}{\partial z}(z(t), p(t), u(t))$
and $\mathcal{H}(z, p, u) = max_{v \in \mathcal{U}} \{\mathcal{H}(z, p, v)\}$.

The maximization condition of (PMP) can be written as $\max_{v \in \mathcal{U}} \{\mathcal{H}(z, p, v)\} = max_{v \in \mathcal{U}} \{\sum_{i=1}^{m} v_i \Phi_i\}$,

where the functions $\Phi_i = pG_i(z)$ are called switching functions. If the controls are bounded, optimal controls are given by bang-bang controls as follows:
if $\Phi_i(t) > 0$ (resp. $\Phi_i(t) < 0$) for $t \in]t_1, t_2[\subset [0, T]$, then the optimal control is bang $u_i(t) = 1$, (resp. $u_i(t) = -1$). If Φ_i is zero over an interval $[t_1, t_2]$, we say that $z(\cdot)$ is a singular extremal trajectory.

For our system (Σ), we write the adjoint vector by bloc $p(s) = (P_1(s), P_2(s))$, with $P_i^t(s) \in R^3$ for $i = 1, 2$ and then $\mathcal{H}(z, p, u) = p_0 + p\dot{z} = p_0 + P_1\dot{x} + P_2\dot{y} =$
$p_0 + P_1((A + u_1 B)x + x \wedge Cy) + P_2((A + u_2 B)y + y \wedge Cx)$.

Proposition 2 The dynamic of the adjoint equation of (Σ) is given by
$\dot{P}_1^t = (A + u_1 B)P_1^t + (P_1^t \wedge Cy) - C(P_2^t \wedge y)$ *and*
$\dot{P}_2^t = (A + u_2 B)P_2^t + (P_2^t \wedge Cx) - C(P_1^t \wedge x)$.

The proof of this proposition uses the following lemma.

Lemma For all $x, y \in R^3$ and for any row vector $q \in R^3$, and for all $(3, 3)$ matrix D with real coefficients one has
1. $x \wedge y = \begin{pmatrix} 0 & y_3 & -y_2 \\ -y_3 & 0 & y_1 \\ y_2 & -y_1 & 0 \end{pmatrix} x$, for $y = (y_1, y_2, y_3)^t$.
2. $\frac{\partial}{\partial x} q(x \wedge y) = -(q^t \wedge y)^t$.
3. $\frac{\partial}{\partial x} q(Dx \wedge y) = -(q^t \wedge y)^t D = -(Dq^t \wedge y)^t$.

The (PMP) maximization condition gives switching functions $\Phi_1 = P_1 Bx$ and $\Phi_2 = P_2 By$. Note that each switching function corresponds to a two-level system. Also, if $\Phi_i(t) > 0$ then $u_i = 1$ and if $\Phi_i(t) < 0$ then $u_i = -1$.

(Σ) is an affine system with two controls that evolves on the sphere of R^6 with radius $\sqrt{2}$. Despite the fact that the Bloch vectors x and y play the same role in the dynamics of (Σ), it is difficult to make an optimal study, since $z \in R^6$ and there is an interaction between x and y which appears in the vector product. Here we consider a particular case to simplify the study of System (Σ). We choose the density matrix $\rho(x, y)$ such that $x = y$ and $u_1 = u_2 = u$. Then the two sub-systems of (Σ) for each Bloch vector follows the same dynamic

$$(\Sigma') \quad \dot{x} = (A + uB)x + x \wedge Cx$$

which represents a system with a Coulomb interaction (when $b \neq c$) and which evolves on the sphere S^2 of R^3. Hamiltonian function is given by

$$\mathcal{H}(x, p, u) = p_0 + p\dot{x} = p_0 + p(A + uB)x + p(x \wedge Cx).$$

The switching function is $\Phi = pBx$ and the adjoint dynamic associated with (Σ') is given by

$$\dot{p^t} = (A + uB)p^t + (p^t \wedge Cx) - C(p^t \wedge x),$$

Note that in this case if in addition $b = c$, then we obtain the same dynamic obtained for the two-level system, and for the state $x(t)$ and for the adjoint vector we have $\dot{x} = (A + uB)x$ *and* $\dot{p^t} = (A + uB)p^t$. But in practice and by definition of the Coulomb parameters, the constants b and c are not necessarily equal.

Theorem For (Σ') $\dot{x} = (A + uB)x + x \wedge Cx$, when $b \neq c$, the coordinate functions $x_i(t)$, $1 \leq i \leq 3$, of the optimal trajectories $x(t)$ associated with $u = \pm 1$ are elliptic functions.

Proof of Theorem We have for a constant control $u = 1$, $\dot{x}_3(t) = x_2(t)$ *and* $\dot{x}_1(t) = ax_2 + (b - c)x_2x_3$. We deduce that $\dot{x}_1(t) = a\dot{x}_3(t) + (b - c)\dot{x}_3x_3$ and hence $x_1(t) = ax_3(t) + \frac{1}{2}(b - c)x_3^2 + k$, *with* $k = -a - \frac{1}{2}(b - c)$, (for $x(0) = (0\,0\,1)^t$). So, $\ddot{x}_3 = \dot{x}_2 = -ax_1 - (b - c)x_1x_3 - x_3$.

We express x_1 as a function of x_3 and we obtain

$$\ddot{x}_3 = \alpha x_3^3 + \beta x_3^2 + \gamma x_3 - ak,$$

with $\alpha = -\frac{1}{2}(b-c)^2$, $\beta = -\frac{3}{2}a(b-c)$ and $\gamma = -a^2 - 1 - (b-c)(a+\frac{1}{2}(b-c))$). It's clear that $\alpha \neq 0$ and $\beta \neq 0$, when $b \neq c$ and $a \neq 0$. It is well-known (see Siegel (1962)) that if $(\alpha, \beta) \neq (0,0)$ then the solutions of this 2-order differential equation, are elliptic and in particular if $(\gamma, ak) = (0,0)$ then the solutions of $\ddot{x}_3 = \alpha x_3^3 + \beta x_3^2$ when $\alpha \neq 0$ are Jacobi elliptic functions.
Notice that $k = 0$ if $2a = c - b$ and then $\gamma = -(a^2+1) \neq 0$ and $\beta = \frac{3}{4}(b-c)^2$ and $\ddot{x}_3 = \alpha x_3^3 + \beta x_3^2 + \gamma x_3$. Recall that Weierstrass elliptic functions are solutions of $\ddot{v} = 4v^3 + g_2 v^2 + g_3 v$, for specific known constants g_2 and g_3.
For the constant control $u = -1$, we find the same coefficients α, β, k and $\ddot{x}_3 = \alpha x_3^3 + \beta x_3^2 + \tilde{\gamma} x_3 - ak$, with $\tilde{\gamma} = -a^2 + 1 - (b-c)(a + \frac{1}{2}(b-c)) = -a^2 + 1 + (b-c)k$. In this case if $k = 0$, $(2a = c - b)$ then $\tilde{\gamma}$ vanishes for $a^2 = 1$ and in this case x_3 is a Jacobi elliptic function.
We deduce that for $u = \pm 1$, x_3 is elliptic and therefore x_2 is elliptic since the derivative of an elliptic function is also elliptic, and x_1 is elliptic since the product and the sum of elliptic functions are elliptic. This proves the theorem.

Recall that for the two-level quantum system, which corresponds to (Σ) when $b = c$, we have found in Zibo et al. (2020) that the coordinate functions $x_i(t)$ for $1 \leq i \leq 3$, of the optimal trajectories $x(t)$ associated with $u = \pm 1$ are trigonometric functions. We can find this result using the above 2-order differential equations with $\alpha = \beta = 0$, $\gamma = -(a^2+1)$, $k = -a$ and $\tilde{\gamma} = -a^2 + 1$.

3. CONCLUSION

This work shows that the geometric study of time minimal control for a two-level quantum system whose dynamic is given by a bilinear control system, which interested many authors, is more simpler than the study of a system with four energy levels. This is due to the presence of several Coulomb physical parameters characterizing the quantum dots and adding nonlinear terms which represent the interaction between energy levels.

REFERENCES

B. Bidgaray-Fesquet and K. Keita, *A nonlinear Bloch model for Coulomb interaction in quantum dots.* J. Math. Phys., 55:021 501, (2014).

B. Bonnard, M. Chyba and D. Sugny, *Time-Minimal Control of Dissipative Two-Level Quantum Systems: The Generic Case.* IEEE Transactions on Automatic Control, Vol. 54, Issue 11, pp. 2598 - 2610, (2009).

U. Boscain and P. Mason, *Time minimal trajectories for two-level quantum systems with drift.* 44th IEEE Conference on Decision and Control, (2005).

U. Boscain and P. Mason, *Time minimal trajectories for a spin 1/2 particle in a magnetic field*, J. Math. Phys., 47: 062 101, (2006).

U. Boscain and B. Piccoli, *Optimal Synthesis for Control Systems on 2-D Manifolds.* Springer, SMAI, 43, (2004).

E. Brning, H. Mkel and A. Messina, *Parameterizations of density matrices.* J. of Modern, 59 : 1-20, (2012).

R.W. Brockett and L. Dai. *Non-holonomic Kinematics and the Role of Elliptic Functions in Constructive Controllability.* Non-holonomic Motion Planning (Z. Li and J. Canny, eds.), Kluwer, Boston, pp. 1-21, (1993).

D. D'Alessandro and M. Dahleh *Optimal Control of Two-Level Quantum Systems.* IEEE Transactions on Automatic Control, Vol. 46, Issue 6, pp. 866 - 876, (2001).

R. El Assoudi-Baikari and A.Gerber, *Optimal Control and Integrability on Lie Groups*, 10th IFAC. NOLCOS 2016, Monterey, California, USA. (49) 18, pp. 994-999, (2016).

R. El Assoudi-Baikari and E. Zibo, *Integrability of geodesics on 4-dimensional sub-Riemannian Lie groups.* 24th International Symposium on MTNS 2020, 54-9, pp. 610 - 614 (2021).

R. Ernst, G. Bodenhausen and A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two dimensions.* Clarendon, Oxford, 1987.

V. Jurdjevic, *Hamiltonian point of view of non-euclidean geometry and elliptic functions.* Systems Control Letters, 43, 25-41, (2001).

N. Khaneja, R. Brockett and S. J. Glaser, *Time optimal control in spin systems*, Phys. Rev. A, 63 (113):131 429 45 50, (2001).

N. Khaneja, S. J. Glaser and R. Brockett, *Sub-riemannian geometry and time optimal control of three spin systems, Quantum gates and coherence transfer.* Phys. Rev. A, 65(3): 032 301 13, (2002).

M. Lapert, E. Assemat, Y. Zhang, S.J. Glaser and D. Sugny, *Time- optimal control of spin 1/2 particles with dissipative and generalized radiation-damping effects.* Phys. Rev. A, 87(4): 043417, (2013).

O.V. Morzhin and A.N. Pechen, *Minimal time generation of density matrices for a two-level quantum system driven by coherent and in coherent controls-International.* J. of Theoretical Phys., Springer, 60: pp. 576-584, (2021).

L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze and E.F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience Publ. John Wiley Sons, New York, (1962).

C.L. Siegel, *Topics in Complex Function Theory I, Elliptic Functions and Uniformization Theory*, volume 1. Addison-Wesley, (1962).

D. Sugny, C. Kontz and H.R. Jauslin, *Time-optimal control of a two- level dissipative quantum system.* Phys. Rev. A, 6(2): 023 419 (2007).

H. Yuan, R. Zeier and N. Khaneja *Elliptic functions and efficient control of Ising spin chains with unequal couplings.* Phys. Rev. A 77, 032340, (2008).

I.E. Zibo *Optimal control : Geometric and numerical methods and applications to the Bloch model, and Integrability of Sub-Riemannian geodesics.* Thesis, I.E. Zibo. LMI, INSA Rouen Normandy (2021).

I.E. Zibo, R. El Assoudi-Baikari, N. Forcadel, *Time Optimal Control of Nonlinear Bloch Equations.* Montes Taurus J. of Pure and Applied Math., 2(2), pp. 49-57, (2020).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Turnpike behaviour for systems that are partially uncontrollable

Martin Gugat * **Martin Lazar** **

* *Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department of Data Science, Lehrstuhl für Dynamics, Control and Numerics (Alexander von Humboldt-Professur), Cauerstr. 11, 91058 Erlangen, Germany (e-mail: martin.gugat@fau.de).*
** *Department of Electrical Engineering and Computing, University of Dubrovnik, Cira Carića 4, 20 000 Dubrovnik, Croatia*

Abstract: We analyse the turnpike properties for a general, linear-quadratic (LQ) optimal control problem. We assume that the system under consideration is governed by an infinite-dimensional differential equation with a generator A of a strongly continuous semi-group. The objective function is the sum of a control cost and a tracking term for an observation of the state.
The novelty of the results is twofold. Firstly, it obtains positive turnpike results for systems that are (partially) uncontrollable. Secondly, it provides turnpike results for optimal averaged control associated to a family of problems that depend on a random parameter, which is the first turnpike type result that extends the averaged controllability approach to optimal control problems. In both cases, the results do not require assumptions on stabilizability and detectability, which are most commonly used in the study of turnpike phenomena.
Examples supporting the theoretical findings will be presented as well.

Keywords: turnpike phenomenon, LQ optimal control problem, infinite-time admissibility, controllability Gramian, observability Gramian.

1. INTRODUCTION

The turnpike property refers to the tendency of optimal controls and trajectories to remain nearly stationary most of the time. It occurs in many optimal control problems associated with a time-evolution system and objective functionals of integral type with a tracking term. It can be interpreted as the property of the optimal control system that the influence of the initial state decays rapidly with time and the same holds for the terminal state backwards in time. This allows a time-dependent control problem to be reduced, at least approximately on the largest part of the time interval, to the corresponding stationary one. Such a simplification is of great interest, both from an application and computational point of view. The term 'turnpike' was coined by economists more than half a century ago and introduced in the context of finite-dimensional, discrete-time optimal control problems. However, it remained out of focus of the mathematical and control community for several decades.

⋆ This work was supported by Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Centre CRC/Transregio 154, Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks, Projects C03 and C05, Projektnummer 239904186.
The research was done while the second author was visiting Chair of Dynamics, Control and Numerics (Alexander von Humboldt Professorship) at Friedrich-Alexander-Universität Erlangen-Nürnberg, with the support of the DAAD (Research Stays for University Academics and Scientists, 2021 programme) and Alexander von Humboldt-Professorship.

Rigorous analysis of the turnpike property started to develop recently in the context of mean field games and model predictive controls (e.g. Cardaliaguet et al. (2012)). A large theory on the topic related to the calculus of variations and optimal control problems has been developed independently by A. Zaslavski in a series of works (cf. Zaslavski (2015) and the references therein). Since then, numerous results have been published in this area, both in finite and infinite dimensional context, as well as for time-discrete and time-continuous systems. The notion has been applied in various contexts: shape design problems, residual neural networks, heat conduction etc. For a detailed introduction we refer an interested reader to some recent, extensive surveys on the topic given in Faulwasser and Grüne (2021); Gershovski and Zuazua (2022).

Most of the results on this topic require the system to be both stabilizable and detectable. In order to obtain the corresponding turnpike properties, the authors analyse the optimality system and explore stabilization properties of the corresponding Riccati operator by using sophisticated functional analysis tools.

Recently, we proposed a new approach based on an estimate for optimal controls that holds in general, without any restrictions on the operators and data that enter the problem (Gugat and Lazar (2021)). One obtains the measure and the integral turnpike property as a direct consequence of the estimate, just by assuming infinite-time admissibility of the control and the observation operator. The results are obtained for systems of both determinis-

tic and stochastic nature. However, the question of the exponential turnpike property, which is stronger than the obtained ones, remained open.

2. PROBLEM SETTING AND THE MAIN RESULT

We consider the control system

$$\begin{aligned} x'(t) + Ax(t) &= Bu(t) \\ x(0) &= x_0, \end{aligned} \tag{1}$$

whose dynamics is governed by an unbounded operator A on a Hilbert space X. Here $u \in L^2_{\text{loc}}([0,\infty);U)$ denotes the control function, U is a Hilbert space, B is a bounded control operator from $\mathcal{L}(U,X)$, and $x^0 \in H$ denotes the initial state.

We consider the optimal control problem

$$\begin{aligned} \min_u J_T(u) = \min_u \frac{1}{2}\int_0^T \left(|u(t)-u_d|_U^2 + |Cx(t)-z_d|_Z^2\right) dt \\ + p_d \cdot y(T), \end{aligned} \tag{2}$$

in which the minimization is taken over the space $L^2_{\text{loc}}([0,\infty);U)$. Here C is an observation operator from $L(X,Z)$, with Z being a Hilbert space, x is the state determined by control u, i.e. it is the solution to (1), u_d and z_d stand for a time independent desirable control and observation, respectively, while $p_d \in X$ determines a linear regularization of the final state.

Using classical convex optimization techniques (e.g. Peypouquet (2015)), the problem (2) is well posed and admits the unique solution given by the formula

$$u_T = -B^* p_T + u_d,$$

where p_T is obtained by solving the corresponding optimality system

$$\begin{aligned} x_T'(t) + Ax_T(t) &= -B(B^*p_T(t) - u_d) \\ x_T(0) &= x_0 \\ -p_T'(t) + A^*p_T(t) &= C^*(Cx_T(t) - z_d) \\ p_T(T) &= p_d. \end{aligned}$$

We also consider the corresponding stationary problem

$$\min_{u\in U} J_s(u) = \min\left\{\frac{1}{2}\left(|u-u_d|_U^2 + |Cx-z_d|_Z^2\right) | Ax = Bu\right\}. \tag{3}$$

Assuming that the stationary state equation $Ax = Bu$ is well posed, the problem (3) admits the unique solution given by

$$\bar{u} = -B^*\bar{p} + u_d,$$

where $\bar{p}$ satisfies the corresponding stationary optimality system

$$A\bar{x} = -B(B^*\bar{p} - u_d) \qquad A^*\bar{p} = C^*(C\bar{x} - z_d),$$

while $\bar{x}$ is the optimal stationary state.

Our examples are motivated by the following result ((Gugat and Lazar, 2021, Theorem 2.1)).

Theorem 1. The difference of solutions to optimal control problems (2) and (3), together with the difference of the corresponding optimal states, satisfies the estimate

$$\begin{aligned} &\|u_T - \bar{u}\|^2_{L^2(0,T;U)} + \|C(x_T - \bar{x})\|^2_{L^2(0,T;Z)} \\ &\le 2\Big(Q_T(x_0-\bar{x})\cdot(x_0-\bar{x}) + \Lambda_T(p_d-\bar{p})\cdot(p_d-\bar{p})\Big), \end{aligned}$$

where Q_T is the observability Grammian for the pair (A,C), while Λ_T stands for the controllability Grammian corresponding to the pair (A,B).

From the last result the basic turnpike properties follow dirrectly ((Gugat and Lazar, 2021, Theorems 2.2-2.3)).

Integral turnpike. The time averages of optimal controls and observations converge strongly:

$$\begin{aligned} \frac{1}{T}\int_0^T u_T &\xrightarrow[T\to\infty]{} \bar{u} \qquad \text{strongly in } U, \\ \frac{1}{T}\int_0^T Cx_T &\xrightarrow[T\to\infty]{} C\bar{x} \qquad \text{strongly in } Z, \end{aligned} \tag{4}$$

with the convergence rate of $O(1/\sqrt{T})$.

Measure turnpike. For every $\varepsilon > 0$ there exists a constant $C_\varepsilon > 0$ such that

$$\mu\big\{t \in [0,T] \big|\ |u_T - \bar{u}|^2 + |C(x_T - \bar{x})|^2 \ge \varepsilon\big\} < C_\varepsilon. \tag{5}$$

Convergence of the (normalized) minimal values.

$$\frac{1}{T}\min J_T \xrightarrow[T\to\infty]{} \min J_s, \tag{6}$$

with the convergence rate of order $1/\sqrt{T}$.

The same kind of results also hold in the stochastic case, i.e. when the dynamics, control and observation operator depend on a random parameter. However, the optimal control is assumed to be parameter independent, which leads to the notion of the optimal averaged control. The turnpike properties (4)-(6) are preserved, with the observation terms being replaced by their averaged values with respect to the parameter, assuming the control and observation operator are infinite-time admissible for almost every value of the random parameter.

REFERENCES

Cardaliaguet, P., Lasry, J.M., Lions, P.L., and Porretta, A. (2012). Long time average of mean field games. *Netw. Heterog. Media*, 7(2), 279–301. doi:10.3934/nhm.2012.7.279. URL `https://doi.org/10.3934/nhm.2012.7.279`.

Faulwasser, T. and Grüne, L. (2021). Turnpike properties in optimal control: An overview of discrete-time and continuous-time results. In E.Z. E. Trélat (ed.), *Numerical Control and Beyond*, volume 22 of *Handbook of Numerical Analysis*, 33 pp. Elsevier.

Gershovski, B. and Zuazua, E. (2022). Turnpike in optimal control of pdes and beyond. *Acta Numerica*, 112 pp.

Gugat, M. and Lazar, M. (2021). Turnpike properties for (partially) uncontrollable systems. *submitted*, 17 pp.

Peypouquet, J. (2015). *Convex optimization in normed spaces: theory, methods and examples*. SpringerBriefs in Optimization. Springer, Cham. doi:10.1007/978-3-319-13710-0.

Zaslavski, A.J. (2015). *Turnpike theory of continuous-time linear optimal control problems*, volume 104 of *Springer Optimization and Its Applications*. Springer, Cham. doi:10.1007/978-3-319-19141-6. URL `https://doi.org/10.1007/978-3-319-19141-6`.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Input Parametrizations and Their Numerical Implementation in Structure-Preserving Optimal Control

Markus Herrmann-Wicklmayr * **Kathrin Flaßkamp** *

* *Systems Modeling & Simulation, Saarland University, Saarbrücken, Germany (e-mail: markus.herrmannwicklmayr@uni-saarland.de)*

Abstract: Motivated by the advantages of structure-preserving integration for applications ranging from molecular dynamics to astrodynamics, geometric integration has been brought into optimal control in the past two decades. Advantages over conventional methods have been shown in biomechanics, robotics, automotive applications, and space mission design. The implicit midpoint method, that is a member of the class of symplectic (partitioned) Runge-Kutta methods but also possess a variational derivation and thus is symmetry-preserving, is widely used due to its many favorable properties. In particular, efficient computations can be achieved by coarse discretizations of state and control signals, since structure preservation does not have to be ensured by small step sizes, as it is the case in conventional methods. Then, specific input parametrizations become an issue when implementing optimized signals in control architectures. We show numerical studies for piecewise linear control signals used in energy optimal control problems.

Keywords: Energy Optimal Control, Structure-Preserving Integration, Implicit Midpoint Method, Input Parametrization, Direct Methods for Optimal Control

1. INTRODUCTION

In the context of structure-preserving (optimal) control the (symplectic) implicit midpoint rule is a popular choice (Ober-Blöbaum et al. (2011), Nair (2012), Leyendecker et al. (2010), Kotyczka and Thoma (2021)) due to its simplicity and many favorable properties (see Section 2). When using the associated Runge-Kutta method to discretize an energy optimal control problem (OCP) in combination with a piecewise linear input parametrization, i.e. a first-order hold (FOH) input signal, undesired oscillatory effects can occur in the computed optimal input trajectory. This extended abstract aims at explaining the origins of the effect and goes into detail on how it can be damped or eliminated. We demonstrate our findings in numerical experiments based on an orbit transition for the Kepler problem.
Notation: We define $[v_1, \dots, v_r] = (v_1^\top, \dots, v_r^\top)^\top$ as the vertical concatenation of the vectors $v_i,\ i = 1, \dots, r$.

2. MOTIVATING THE IMPLICIT MIDPOINT METHOD

The implicit Runge-Kutta method known as *implicit midpoint method (IMP)* (with the Butcher Tableau coefficients $(\mathsf{A}, \mathsf{b}, \mathsf{c}) = (\frac{1}{2}, 1, \frac{1}{2})$) has the following properties. It is of second order, has $s = 1$ stages, is A-stable, symmetric and symplectic. A special feature is that the implicit stage equation can be eliminated for any nonlinear system dynamics. This results in the implicit difference equation

$$x_{k+1} = x_k + hf\left(t_{k+\frac{1}{2}}, \frac{x_{k+1} + x_k}{2}\right), \quad t_{k+\frac{1}{2}} := \left(k + \tfrac{1}{2}\right)h,$$

where h is the step size. Then, for a full discretization of the OCP, the IMP requires just as many optimization variables (OVs) as an explicit Runge-Kutta method.

The A-stability of the method allows larger step sizes where explicit methods would become numerically unstable. By this, the size of the optimization problem can be kept small. This typically results in shorter optimization times. The before-mentioned properties are particularly suitable for real-time optimal control schemes, like model predictive control, or if a solution on a coarse time grid is required as a warmstart to an OCP on a finer time grid.

3. MOTIVATING FIRST-ORDER HOLD CONTROL

When solving an OCP for a non-autonomous continuous time (CT) system $\dot{x}(t) = f(t, x(t), u(t)),\ x \in \mathbb{R}^n,\ u \in \mathbb{R}^m$ via a direct method one needs to use a discrete time (DT) approximation of the CT dynamics. When using a s-stage Runge-Kutta method the time dependent parts of the vector field $f(t, x(t), u(t))$ are evaluated at s points to compute the next state. This requires the knowledge of $u(t)$ over the whole prediction horizon. Since the infinite-dimensional CT OCP needs to be approximated by a finite-dimensional optimization problem, the number of variables that parametrize the input has to be limited. The most common approach is to define the input to be constant on the time interval between two time instances. Then, for N time instances there are at most $N - 1$ inputs or $(N - 1) \cdot m$ variables that parametrize the input trajectory. This easy-to-implement approach is known as zero-order hold (ZOH) and it comes with some drawbacks. The ZOH approach

a) implicitly assumes that the input of the plant system, that is to be controlled, can jump between two values,
b) can lead to a large overshoot / peak in the states when applied to the plants (Herrmann-Wicklmayr et al. (2022)),
c) and thus would require a very fine time grid to capture this behavior.

These points mean that for a sufficiently large step size and/or *jump height* the behavior of the DT model and the CT plant might differ significantly. Consequently, the computed costs do not predict the true costs sufficiently accurate, the tracking performance might be poor, and unpredicted damaging of systems components might occur.

One now could argue that the above-mentioned problems can be circumvented by linearly connecting the computed inputs of the ZOH OCP. However, this would mean that the input trajectory implicitly assumed within the OCP, $u_{\mathrm{OCP}}(t) = u_k = const.$, $t \in [t_k; t_{k+1}) \coloneqq \mathcal{T}_k$, and the actually applied input trajectory

$$u_{\mathrm{plant}}(t) = u_k + \frac{u_{k+1} - u_k}{t_{k+1} - t_k}(t - t_k),\ t \in \mathcal{T}_k$$

differ (unless $u_{k+1} = u_k$). The CT trajectory error is

$$e_{\Delta u,\mathrm{CT}}(k) = \int_{t_k}^{t_{k+1}} |u_{\mathrm{plant}}(t) - u_{\mathrm{OCP}}(t)|\,\mathrm{d}t = \tfrac{h_k}{2}|u_{k+1} - u_k|$$

and it grows linearly with the step size $h_k = t_{k+1} - t_k$ and with the input difference. It only vanishes for $h_k \to 0$ and/or $|u_{k+1} - u_k| \to 0$. The DT input trajectory error is

$$e_{\Delta u,\mathrm{DT}}(k) = \sum_{i=1}^{s} \left|u_{\mathrm{plant}}(t_{k,i}) - u_{\mathrm{OCP}}(t_{k,i})\right| = \sum_{i=1}^{s} c_i\,|u_{k+1} - u_k|,$$

with $t_{k,i} = t_k + c_i h_k$. The error is only guaranteed to be zero if $\sum_{i=1}^{s} c_i = 0$. However, that is solely fulfilled by the first-order explicit Euler method. In summary, this implies that a linear interpolation only avoids imposing additional errors, if the explicit Euler method is used for the discretization.

For these reasons, the next logical step is to directly consider a FOH input trajectory within the OCP, i.e.

$$\tilde{u}(t) = u_k + \frac{u_{k+1} - u_k}{t_{k+1} - t_k}(t - t_k), \quad t \in \mathcal{T}_k$$

defined on a not necessarily equidistant time grid $t_0, \dots, t_N$ and via the nodes $u_0, \dots, u_N$. This approach comes at the negligible cost of m more OVs, representing the input at the final time instance, and possibly an m-dimensional equality constraint, expressing the initial condition of the input, and a slightly more involved implementation. Although not necessary, it might be useful to constrain the inputs rate of change in additional inequality constraints.

4. OPTIMAL INPUT TRAJECTORIES WITH ARBITRARY LARGE OSCILLATIONS

Consider the CT OCP

$$\begin{aligned} \min_{u(\cdot)}\ & J\big(u(\cdot)\big) = \int_0^T \ell(t)\,\mathrm{d}t \\ \text{s.t.}\ & \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x^0, \\ & x(t) \in \mathbb{X}\ \forall\ t \in [0; T], \quad x(T) = x^T, \end{aligned} \qquad \left(\mathfrak{P}^{\mathrm{CT}}\right)$$

with the running cost $\ell(t) \coloneqq u(t)^\top R u(t)$ and the state constraint set $\mathbb{X}$. W.l.o.g. we assume that $R = I_m$. This is justified since with any positive definite matrix and its decomposition $\hat{R} = B^\top B$ (B is invertible) we can write $\hat{u}^\top \hat{R} \hat{u} = \left(B\hat{u}\right)^\top B\hat{u} = u^\top I_m u$ with $u = B\hat{u}$.

Let $\varphi\left(x^0, 0, t, u|_{[0;t]}\right) = x(t)$ be the flow of the CT system starting at $x(0)$ under the control $u|_{[0;t]}$ that is, in general, not available in closed form. Then, with $\tilde{\varphi}\left(x^0, 0, k, u|_{[0;t_k]}\right) = x_k$ we denote a DT approximation of the flow that depends on the chosen Runge-Kutta method. We apply the same method to the cost functional and obtain the approximation

$$\tilde{J} = \sum_{k=0}^{N-1} \sum_{i=1}^{s} b_i h \ell(t_k + c_i h) = \sum_{i=1}^{s} b_i \sum_{k=0}^{N-1} h u_{k,i}^\top u_{k,i}$$

with $u_{k,i} = u(t_k + c_i h)$. Using the IMP and defining $u_{k,1}$ as $u_{k+\frac{1}{2}} = \tilde{u}(t_k + \frac{1}{2}h)$ we obtain the DT OCP

$$\begin{aligned} \min_{U}\ & \tilde{J}(U) = h \sum_{k=0}^{N-1} u_{k+\frac{1}{2}}^\top u_{k+\frac{1}{2}} \\ \text{s.t.}\ & x_{k+1} = \tilde{\varphi}_{\mathrm{IMP}}\left(x_k, k, k+1, \tilde{u}|_{\mathcal{T}_k}\right), \quad x_0 = x^0, \\ & x_k \in \mathbb{X}\ \forall\ k \in \mathbb{N}_{1,N-1}, \quad x_N = x^T, \\ & \tilde{u}(t) = u_k + \tfrac{u_{k+1} - u_k}{h}(t - t_k), \quad t \in \mathcal{T}_k \end{aligned} \qquad \left(\mathfrak{P}^{\mathrm{IMP}}\right)$$

with $U = [u_0, u_1, \dots, u_N]$.

Remark. *As $h \to 0$, the solution of $\mathfrak{P}^{\mathrm{IMP}}$ does not necessarily converge to the one of $\mathfrak{P}^{\mathrm{CT}}$. This is caused by the non-coercivity of the discretized cost summands*

$$u_{k+\frac{1}{2}}^\top u_{k+\frac{1}{2}} = \tfrac{1}{4}(u_{k+1} + u_k)^\top (u_{k+1} + u_k)$$

as pointed out in Campos et al. (2015). This is due to the combination of the chosen Runge-Kutta scheme and input parametrization.

Proposition 4.1. *Assume that U^* is a local optimal solution to $\mathfrak{P}^{\mathrm{IMP}}$ and results in the cost $\tilde{J}^* = \tilde{J}\left(U^*\right)$. Then, for every optimal solution U^*, there exist infinitely many optimal solutions $\check{U}^*(A)$, $A \in \mathbb{R}^m$ that result in the same cost $\tilde{J}^* = \tilde{J}\left(\check{U}^*(A)\right)$, $i \in \mathbb{N}$.*

Proof. We define $\tilde{u}^*(\cdot)$ to be the optimal trajectory fully determined by U^*. Now $\tilde{u}^*(\cdot)$ is additively perturbed by

$$\Delta u(t, A) = A(-1)^k \left(1 - \tfrac{2}{h}(t - t_k)\right), \quad t \in \mathcal{T}_k$$

resulting in $u^d(t, A) = \tilde{u}^*(t) + \Delta u(t, A)$. The function $u^d(t, A)$ is piecewise linear by construction and has the property that $u^d(t_{k+\frac{1}{2}}, A) = \tilde{u}^*(t_{k+\frac{1}{2}})\ \forall k \in \mathbb{N}_{0,N},\ \forall A \in \mathbb{R}^m$. Equivalently, one can say that with $u_k^d = u_k^* + A(-1)^k$ the equality

$$u_{k+1}^d + u_k^d = u_{k+1}^* + A(-1)^{k+1} + u_k^* + A(-1)^k = u_{k+1}^* + u_k^* \quad (2)$$

holds. Then, $U^d = U^* + [A, -A, \dots, A \cdot (-1)^N]$ and the resulting state sequences $X^d = [x_0^d, \dots, x_N^d]$ are the same for U^* and U^d, i.e. $X^d = X^*$. Hence, $Z^d = [U^d, X^*]$ is a feasible solution to the nonlinear program (NLP) detailed out below.

Next, we show the optimality of the solution. We assume that $x_k \in \mathbb{X}$ can be expressed via an inequality and define

$$\begin{aligned} c_k &= c(x_k, x_{k+1}, u_k, u_{k+1}) \\ &= x_{k+1} - x_k - h f\left(t_{k+\frac{1}{2}}, \tfrac{x_{k+1} + x_k}{2}, \tfrac{u_{k+1} + u_k}{2}\right) \end{aligned}$$

for all $k \in \mathbb{N}_{0,N-1}$, representing the continuity constraint of the states. Then, with $Z = [U, X]$ the function $\hbar(Z)$

and vector λ represents all equality constraints (continuity constraints & initial/terminal constraint) and their corresponding Lagrange multipliers, the function $g(Z)$ and vector μ all inequality constraints (state constraints) and their corresponding Lagrange multipliers in the NLP derived from $\mathfrak{P}^{\mathrm{IMP}}$.

Then, $\mathfrak{P}^{\mathrm{IMP}}$ can be written as a nonlinear optimization problem or NLP of the form

$$\min \tilde{J}(Z) \;\text{ s.t. } \hbar(Z) = 0,\; g(Z) \leq 0$$

with $Z = [U, X]$. Since U^* was an optimal solution to $\mathfrak{P}^{\mathrm{IMP}}$ the DT dynamics output an optimal state sequence X^* and we know that there exist Lagrange multipliers λ^*, μ^* such that the triple (Z^*, λ^*, μ^*) satisfies all Karush-Kuhn-Tucker (KKT) conditions.

Based on the definition of $u_{k+\frac{1}{2}}$ and the chosen input parametrization we can rewrite the cost functional as

$$\tilde{J}(Z) = \tilde{J}(U) = \frac{h}{4} \sum_{k=0}^{N-1} u_{k+1}^\top u_{k+1} + 2u_{k+1}^\top u_k + u_k^\top u_k.$$

Then the Lagrangian of the NLP is

$$L(Z, \lambda, \mu) = \tilde{J}(Z) + \lambda^\top \hbar(Z) + \mu^\top g(Z).$$

With the placeholder $(*) := \left(t_{k+\frac{1}{2}}, \frac{\bar{x}_{k+1}+\bar{x}_k}{2}, \frac{\bar{u}_{k+1}+\bar{u}_k}{2}\right)$, it follows that

$$\begin{aligned} \nabla_{x_k} c_k\big|_{(*)} &= -I - \tfrac{h}{2}\, \nabla_{x_k} f\big|_{(*)}, \quad \nabla_{u_k} c_k\big|_{(*)} = -\tfrac{h}{2}\, \nabla_{u_k} f\big|_{(*)}, \\ \nabla_{x_{k+1}} c_k\big|_{(*)} &= I - \tfrac{h}{2}\, \nabla_{x_{k+1}} f\big|_{(*)}, \end{aligned} \tag{3}$$

for all $k \in \mathbb{N}_{0,N-1}$ and $k \in \mathbb{N}_{0,N}$, respectively. In the same manner we obtain

$$\begin{aligned} \nabla_{u_k} \tilde{J}(Z)\big|_{\bar{Z}} &= h\left(\bar{u}_k + \tfrac{1}{2}\left(\bar{u}_{k-1} + \bar{u}_{k+1}\right)\right) \quad \forall\; k \in \mathbb{N}_{1,N-1}, \\ \nabla_{u_0} \tilde{J}(Z)\big|_{\bar{Z}} &= \frac{1}{2} h \cdot (\bar{u}_1 + \bar{u}_0), \\ \nabla_{u_N} \tilde{J}(Z)\big|_{\bar{Z}} &= \frac{1}{2} h \cdot (\bar{u}_N + \bar{u}_{N-1}). \end{aligned} \tag{4}$$

With a slightly more compact notation, the Lagrange condition is

$$\nabla_Z L(Z^*, \lambda^*, \mu^*) = \nabla_Z \tilde{J}(Z^*) + \nabla_Z \hbar(Z^*)\lambda^* + \underbrace{\nabla_Z g(Z^*)}_{=\nabla_Z g(X^*)} \mu^* \overset{!}{=} 0.$$

We refrain from explicitly writing down the primal feasibility, dual feasibility and complementary slackness conditions (remaining KKT conditions). Now set the dynamically feasible choice of OVs $\bar{Z} = Z^d = [U^d, X^*]$. From (2) and the definition of f we can conclude that (3) and (4) evaluated at $(t_{k+\frac{1}{2}}, x_k^*, x_{k+1}^*, u_k^d, u_{k+1}^d)$ or $(t_{k+\frac{1}{2}}, x_k^*, x_{k+1}^*, u_k^*, u_{k+1}^*)$ yield the same results and it follows

$$\nabla_Z Y\left(Z^*\right) = \nabla_Z Y\left(Z^d\right), \quad Y = \{\tilde{J}, \hbar, g\}.$$

Hence, if the Lagrange condition is fulfilled for (Z^*, λ^*, μ^*), then it is fulfilled for $(Z^d, \lambda^d, \mu^d) = ([U^d, X^*], \lambda^*, \mu^*)$. The same holds true for the remaining KKT conditions. Moreover, all previous arguments are valid for any $A \in \mathbb{R}^m$ and $\tilde{U}^*(A) = U^d$. □

Next, among all the possible input trajectories $\tilde{u}^{\bigstar}(t) = \tilde{u}^*(t) + \Delta u(t, A)$, $A \in \mathbb{R}^m$, where $\tilde{u}^*(t)$ is any known (locally) optimal solution to $\mathfrak{P}^{\mathrm{IMP}}$, we determine the one optimizing the *actual* cost $J^{\bigstar} = \int_0^T \left(\tilde{u}^{\bigstar}(t)\right)^\top \tilde{u}^{\bigstar}(t)\,\mathrm{d}t$ from $\mathfrak{P}^{\mathrm{CT}}$. The choice of the parameter A neither affects the cost nor the state trajectory (in $\mathfrak{P}^{\mathrm{IMP}}$). We now determine A such that $J^{\bigstar}$ is minimized. First we compute the cost

$$\begin{aligned} J^{\bigstar} &= \int_0^T \left(\tilde{u}^{\bigstar}\right)^\top \tilde{u}^{\bigstar}\,\mathrm{d}t \\ &= \int_0^T \left(\tilde{u}^*\right)^\top \tilde{u}^* + 2\left(u^*\right)^\top \Delta u + \Delta u^\top \Delta u\,\mathrm{d}t \\ &= \ldots = \frac{h}{3}\left(N \cdot A^\top A + \sum_{k=0}^{N-1} (-1)^k A^\top \left(u_k^* - u_{k+1}^*\right)\right) + J^*. \end{aligned}$$

Finding the extremum of $J^{\bigstar}$ yields the optimal parameter choice

$$\frac{\partial J^{\bigstar}}{\partial A}\left(A^{\bigstar}\right) \overset{!}{=} 0 \;\Leftrightarrow\; A^{\bigstar} = -\frac{1}{2N} \sum_{k=0}^{N-1} (-1)^k \left(u_k^* - u_{k+1}^*\right). \tag{5}$$

This resolves the ambiguity of the DT OCP based on the IMP by adding a *post-processing* step.

Besides this post-processing ansatz, there are (at least) two approaches on how to prevent the oscillatory behavior:
a) Directly use the analytical solution of the cost functional. Consider the piecewise linear function $z(t) \in \mathbb{R}^m$ that is defined as $\tilde{u}(t)$ in $\mathfrak{P}^{\mathrm{IMP}}$. Then we can easily compute the integral

$$J = \int_0^T z(t)^\top z(t)\,\mathrm{d}t = \frac{h}{3} \sum_{k=0}^{N-1} z_{k+1}^\top z_{k+1} + z_k^\top z_{k+1} + z_k^\top z_k.$$

b) Use an integrator that *evaluates* the input trajectory $\tilde{u}(t)$ at *multiple* points $t_{k,i} := t_k + c_i h$, $i = 1, \ldots, s$ and weighs these points with (optimally) non-vanishing coefficients $b_i > 0$:

$$\tilde{J}(U) = \sum_{k=0}^{N-1} \sum_{i=1}^{s} b_i h \ell(t_k + c_i h) = \sum_{i=1}^{s} b_i \sum_{k=0}^{N-1} h u_{k,i}^\top u_{k,i} =: \sum_{i=1}^{s} b_i \tilde{J}_{u,i} \tag{6}$$

with $u_{k,i} := u(t_k + c_i h)$. At least one of the points should be *sufficiently far* away from the midpoint and have a *sufficiently large* weight, i.e. there exists at least one coefficient pair (b_i, c_i) that satisfies $|c_i - \frac{1}{2}| \cdot |b_i| > \Delta_{\mathrm{MP}} > 0$. A priori it is not clear what *sufficiently far* and *large* means, i.e. how large Δ_{MP} has to be. Assuming the inequality is satisfied, arbitrary large oscillations in the input trajectory would be captured and, in conclusion, be damped or eliminated as a consequence of the optimization process.

As an example, consider the generic second order Runge-Kutta method with the Butcher tableau

$$\begin{array}{c|c} \mathsf{c} & \mathsf{A} \\ \hline & \mathsf{b} \end{array} \;\hat{=}\; \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \alpha_1 & \alpha_2 & 0 \\ \hline & \delta_{\mathrm{MP}} & \delta_{\mathrm{MP}} & 1 - 2\delta_{\mathrm{MP}} \end{array} \;\hat{=}\; \mathrm{BT}^{\mathrm{EMP}_\delta}$$

with $\delta_{\mathrm{MP}} \in [0; \frac{1}{2}]$ and $\alpha_1 + \alpha_2 = \frac{1}{2}$. We set $\alpha_2 = 0$ which results in $\alpha_1 = \frac{1}{2}$. For $\delta_{\mathrm{MP}} = 0$ we obtain the explicit midpoint method (EMP). Otherwise for $i = 1, 2$ we have $|c_i - \frac{1}{2}| \cdot |b_i| = \frac{1}{2}\delta_{\mathrm{MP}}$. Using the parameter δ_{MP}, this allows us to control how much the method, which we call EMP_δ, deviates from the EMP. We replace the cost computation in $\mathfrak{P}^{\mathrm{IMP}}$ with (6) and with the coefficients of $\mathrm{BT}^{\mathrm{EMP}_\delta}$. We denote the resulting OCP as $\mathfrak{P}^{\mathrm{IMP,EMP}_\delta}$.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

5. NUMERICAL RESULTS

For our numerical investigation we use the Kepler problem dynamics (in Hamilton formulation)

$$\dot{x} = \frac{\mathrm{d}}{\mathrm{d}t}\left[r,\ \theta,\ p_r,\ p_\theta\right] = \left[\tfrac{p_r}{m},\ \tfrac{p_\theta}{mr^2},\ \tfrac{1}{mr^3}p_\theta^2 - \tfrac{\gamma mM}{r^2} + u_r,\ u_\theta\right]$$

with parameters $k = \gamma mM = 1 \cdot 10^3$ and $m = 1$. An energy-optimal (in the sense of $\mathfrak{P}^{\mathrm{CT}}$) transition from one circular orbit $x_s(r,\theta_s) = \left[r_s, \theta_s, 0, mr_s^2\sqrt{\frac{k}{mr_s^3}}\right]$ to another one should be finished in $T = 3\,\mathrm{s}$. We set the initial state condition to $x^0 = x_s(5,0)$, the terminal one to $x^T = x_s(6,\theta^T)$ with $\theta^T \in \mathbb{R}$, i.e. θ^T is unconstrained. We discretize the CT OCP by using the IMP for both dynamics and running cost. If additionally an FOH input parametrization is assumed, we obtain a problem of the type $\mathfrak{P}^{\mathrm{IMP}}$.

We solve the corresponding NLP for multiple numbers of shooting nodes and compare the resulting input trajectory with a reference solution $u_{\mathrm{ref}} = [u_{r,\mathrm{ref}}, u_{\theta,\mathrm{ref}}]$ (obtained from solving a DT OCP using the classical RK4 scheme and a step size $h = 10^{-3}\,\mathrm{s}$). The difference of the trajectories is evaluated at the nodes, i.e. $e_k = u_k - u_{\mathrm{ref}}(t_k)$, and the mean square error (MSE) is computed. The results are shown in Fig. 1. No systematic change in the MSE with increasing N, i.e. decreasing h, can be observed. This is plausible since, as it was shown in Section 4, all oscillation amplitudes are – in theory – equally likely. The obtained solution *highly* depends on the chosen hyperparameters of the solver and numerical noise. This behavior is not observed in general.

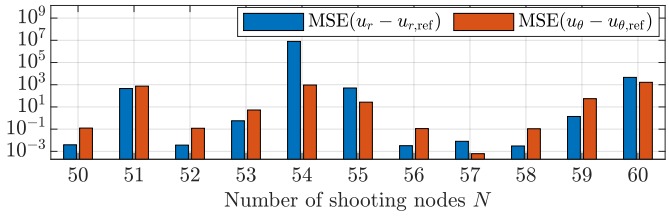


Fig. 1. The OCP $\mathfrak{P}^{\mathrm{IMP}}$ is solved for different number of shooting nodes, resulting in different step sizes $h = \frac{T}{N}$ that range from 0.05 s to 0.06 s.

Next, we show the effect of post-processing the previously found solutions by computing $A^\star$ and adding $u^d(t, A^\star)$ to them. This results in Fig. 2. The errors are in the same order and a trend – the error decreases as N increases – can be observed.

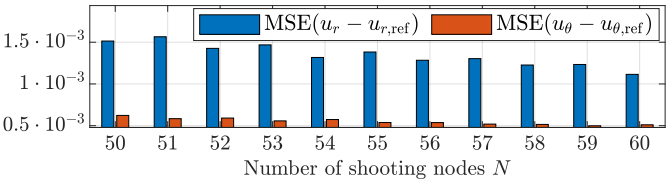


Fig. 2. The solution used to generate Fig. 1 were post-processed by adding $\Delta u(t, A^\star)$, where $A^\star$ is computed as shown in (5). The y-axis is non-logarithmic.

Our next investigation concerns the effect of using $\mathfrak{P}^{\mathrm{IMP,EMP}_\delta}$ with different parameters δ_{MP}, i.e. a different (generic) Runge-Kutta method with the Butcher tableau $\mathrm{BT}^{\mathrm{EMP}_\delta}$ was used to integrate the running costs. We choose $N = 60$ shooting nodes which corresponds to one of the worst solutions of Fig. 1. The results are shown in Fig. 3. Even a small weight $b_1 = b_2 = \delta_{\mathrm{MP}} \geq 4 \cdot 10^{-4}$ on the cost parts $\tilde{J}_{u,i}$, see (6), results in a MSE smaller than 10^{-4} for both inputs u_r, u_θ. It is clearly visible that the larger the deviation δ_{MP} becomes, the smaller the MSE of the inputs. This confirms the intuition that the sensitivity of the solution is a result of the EMP/IMP, rather than the chosen hyperparameters of the solver.

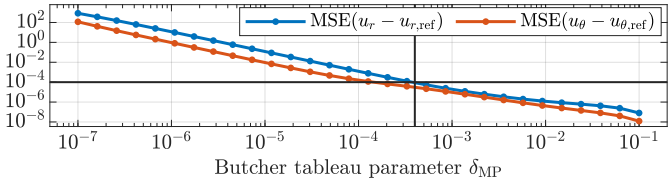


Fig. 3. MSEs when solving the $\mathfrak{P}^{\mathrm{IMP,EMP}_\delta}$ for various δ_{MP}.

6. CONCLUSION

In this extended abstract we showed that the popular choice of the implicit midpoint method (in the context of energy optimal control) can lead to undesired and unpredictable numerical effects in the form of arbitrarily large oscillations. Thereby one possible effect of the non-coercivity, as described in Campos et al. (2015), was presented. We then demonstrate how the oscillations can be damped or eliminated a) by post-processing the OCP solution and b) by removing the non-coercivity property. We conjecture that similar effects for other types of problems, e.g. discretizing holonomic constraints (Johnson and Murphey (2009)), result from applying the IMP.

REFERENCES

Campos, C.M., Ober-Blöbaum, S., and Trélat, E. (2015). High order variational integrators in the optimal control of mechanical systems. *Discrete & Continuous Dynamical Systems - A*, 35(9), 4193–4223.

Herrmann-Wicklmayr, M., Rizzello, G., and Flaßkamp, K. (2022). Numerically Efficient Discrete-Time Dielectric Elastomer Actuators Models for Optimal Control. In *2022 Conference on Mathematical Modelling (accepted)*.

Johnson, E.R. and Murphey, T.D. (2009). Dangers of two-point holonomic constraints for variational integrators. In *2009 American Control Conference*, 4723–4728.

Kotyczka, P. and Thoma, T. (2021). Symplectic discrete-time energy-based control for nonlinear mechanical systems. *Automatica*, 133, 109842.

Leyendecker, S., Ober-Blöbaum, S., Marsden, J.E., and Ortiz, M. (2010). Discrete mechanics and optimal control for constrained systems. *Optimal Control Applications and Methods*, 31(6), 505–528.

Nair, S. (2012). Time adaptive variational integrators: A space–time geodesic approach. *Physica D: Nonlinear Phenomena*, 241(4), 315–325.

Ober-Blöbaum, S., Junge, O., and Marsden, J.E. (2011). Discrete mechanics and optimal control: An analysis. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(2), 322–352.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Recent Advances in Decoding General Lee Metric Codes

Jessica Bariffi [*,**] **Karan Khathuria** [***,⋆] **Violetta Weger** [****]

[*] *Institute of Communication and Navigation, German Aerospace Center, Germany (e-mail: jessica.bariffi@dlr.de)*
[**] *Institute of Mathematics, University of Zurich, Switzerland*
[***] *Institute of Computer Science, University of Tartu, Estonia (e-mail: karan.khathuria@ut.ee).*
[****] *Department of Electrical and Computer Engineering, Technical University of Munich, Germany (e-mail: violetta.weger@tum.de)*

Abstract: The prospect and demand of using the Lee metric to construct public-key encryption schemes and digital signature schemes have immensely grown in recent times. This leads the researchers to ask about the hardness of decoding a general Lee metric code. In this work, we answer this question by showing that the syndrome decoding problem over the Lee metric is NP-complete. Moreover, we will quantify the computational hardness of the syndrome decoding problem with respect to the best-known ISD algorithms.

Keywords: Lee metric, Information-set decoding, Syndrome decoding problem
2010 MSC: 11T71, 94B35

1. INTRODUCTION

The problem of decoding a general linear code has recently gained immense interest due to its cryptographic applications. McEliece, in [11], proposed a public-key encryption scheme whose security is based on the hardness of decoding a general linear code over the Hamming metric. Currently, this cryptosystem stands as one of the most promising candidates for post-quantum cryptography [6]. However, McEliece's cryptosystem comes with one drawback of having large key sizes. This has encouraged many researchers to look for an alternative way to create code-based cryptosystems. One such way is to change the underlying metric, such as to rank metric or Lee metric.

In this work, we study the hardness of decoding a general linear code over the Lee metric. We first show that the Lee metric syndrome decoding problem (L-SDP) is NP-complete, which follows a similar proof as in the case of the Hamming metric. Next, we discuss the computational hardness of solving the L-SDP problem for a random instance case. In particular, we discuss all the recently developed decoding algorithms and compare their performance. We further provide some new ideas to improve the existing algorithms and develop new algorithms for L-SDP.

2. PRELIMINARIES

In this section, we present some preliminaries about the Lee metric and linear codes over an integer residue ring. Throughout this section, let m, n be positive integers.

Definition 1. (1) For $x \in \mathbb{Z}/m\mathbb{Z}$, the *Lee weight* of x is given by

$$\mathrm{wt}_L(x) := \min\{x, |m-x|\}.$$

(2) For a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in (\mathbb{Z}/m\mathbb{Z})^n$, the *Lee weight* of $\mathbf{x}$ is given by

$$\mathrm{wt}_L(\mathbf{x}) := \sum_{i=1}^{n} \mathrm{wt}_L(x_i).$$

Definition 2. A *linear code* $\mathcal{C}$ of length n over $\mathbb{Z}/m\mathbb{Z}$ is a $\mathbb{Z}/m\mathbb{Z}$-submodule of $(\mathbb{Z}/m\mathbb{Z})^n$.

Let $m = p_1^{s_1} p_2^{s_2} \cdots p_\ell^{s_\ell}$ be the prime factorization of m. Then, by Chinese remainder theorem, we know that $\mathbb{Z}/m\mathbb{Z} \cong (\mathbb{Z}/p_1^{s_1}\mathbb{Z}) \times (\mathbb{Z}/p_2^{s_2}\mathbb{Z}) \times \cdots \times (\mathbb{Z}/p_\ell^{s_\ell}\mathbb{Z})$. Moreover, $\mathcal{C} \cong \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_\ell$, where $\mathcal{C}_i$ is a linear code over $\mathbb{Z}/p_i^{s_i}\mathbb{Z}$ for each $i = 1, \ldots, \ell$. Hence, we may restrict our study to $m = p^s$, for some prime p and integer s.

Definition 3. Let $\mathcal{C}$ be a linear code of length n over $\mathbb{Z}/p^s\mathbb{Z}$. Then,

(1) The *type* of $\mathcal{C}$ is the partition $\lambda = (\lambda_1, \ldots, \lambda_K)$ such that
$$\mathcal{C} \cong (\mathbb{Z}/p^{\lambda_1}\mathbb{Z}) \times (\mathbb{Z}/p^{\lambda_2}\mathbb{Z}) \times \cdots \times (\mathbb{Z}/p^{\lambda_K}\mathbb{Z}),$$
with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K \geq 0$. The type λ can also be denoted by $(s^{k_1}, (s-1)^{k_2}, \ldots, 1^{k_s})$, where k_i is the multiplicity of the part of size $s-i+1$ in λ.
(2) The number of parts K of λ is called the *rank* of $\mathcal{C}$.
(3) A *generator matrix* of $\mathcal{C}$ is a matrix $\mathbf{G}$ over $\mathbb{Z}/p^s\mathbb{Z}$ such that its rows generate $\mathcal{C}$.
(4) A *parity-check matrix* of $\mathcal{C}$ is a matrix $\mathbf{H}$ over $\mathbb{Z}/p^s\mathbb{Z}$ such that its nullspace is $\mathcal{C}$.
(5) An *information set* is a subset I of $\{1, 2, \ldots, n\}$ having minimal size such that
$$|\{\mathbf{c}_I : \mathbf{c} \in \mathcal{C}\}| = |\mathcal{C}|,$$
where $\mathbf{c}_I$ is a vector with entries from $\mathbf{c}$ that are indexed in I.

⋆ Corresponding author.

For a code $\mathcal{C}$ of rank K, it is easy to see that the cardinality of an information set is K.

3. LEE SYNDROME DECODING PROBLEM

Let n, m, k be positive integers with $k \leq n$. Then the syndrome decoding problem over the Lee metric is defined as follows.

Problem 1. (Lee syndrome decoding problem (L-SDP)). Given a parity-check matrix $\mathbf{H} \in (\mathbb{Z}/m\mathbb{Z})^{(n-k)\times n}$, a syndrome $\mathbf{s} \in (\mathbb{Z}/m\mathbb{Z})^{n-k}$ and a positive integer t, find $\mathbf{e} \in (\mathbb{Z}/m\mathbb{Z})^n$ such that $\mathrm{wt}_L(\mathbf{e}) \leq t$ and $\mathbf{e}\mathbf{H}^\top = \mathbf{s}$.

In [14, Proposition 2], it is proved that L-SDP is NP-complete, by reducing the 3-dimensional matching (3DM) problem to SDP. The proof is very similar to the proof of NP-completeness of SDP for the Hamming metric binary codes [4] and non-binary codes [1].

4. LEE INFORMATION SET DECODING ALGORITHMS

Many of the information set decoding (ISD) algorithms over the Hamming metric have been adapted to the Lee metric. In this section, we discuss some these adaptations.

Two-blocks algorithm It is an adaptation of Stern's Hamming ISD algorithm [12] for the Lee metric codes over the ring $\mathbb{Z}/p^s\mathbb{Z}$. This algorithm also generalizes the Lee-Brickell's and Prange's ISD algorithms.

The idea of Stern's algorithm is to partition the chosen information set I into two sets X and Y containing v_1 and v_2 errors, respectively. Moreover, it is assumed that there exists a zero-window Z of size z outside of the information set where no errors happen. For a complete description of the algorithm, see [14, Algorithm 1].

s-blocks algorithm In this algorithm, we use the structure of the ring $\mathbb{Z}/p^s\mathbb{Z}$. For a linear code $\mathcal{C}$ over $\mathbb{Z}/p^s\mathbb{Z}$ of length n and type $(s^{k_1}, \ldots, 1^{k_s})$, a parity-check matrix can be written in the following systematic form (up to permutation of columns):

$$\mathbf{H} = \begin{pmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \cdots & \mathbf{A}_{1,s} & \mathrm{Id}_{n-K} \\ p\mathbf{A}_{2,1} & p\mathbf{A}_{2,2} & \cdots & p\mathrm{Id}_{k_s} & \mathbf{0}_{k_s\times(n-K)} \\ p^2\mathbf{A}_{3,1} & p^2\mathbf{A}_{3,2} & \cdots & \mathbf{0}_{k_{s-1}\times k_s} & \mathbf{0}_{k_{s-1}\times(n-K)} \\ \vdots & \vdots & & \vdots & \vdots \\ p^{s-1}\mathbf{A}_{s,1} & p^{s-1}\mathrm{Id}_{k_2} & \cdots & \mathbf{0}_{k_2\times k_s} & \mathbf{0}_{k_2\times(n-K)} \end{pmatrix}$$

Using this structure, we split the error vector into $s+1$ parts, i.e., $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_{s+1})$ with $\mathbf{e}_i \in (\mathbb{Z}/p^s\mathbb{Z})^{k_i}$ for $1 \leq i \leq s$ and $\mathbf{e}_{s+1} \in (\mathbb{Z}/p^s\mathbb{Z})^{n-K}$. Next, we assume a fixed weight distribution on the $s+1$ parts of the error vector, and find the error vector satisfying all the s equations obtained using the systematic form.

Partial Gaussian elimination algorithms In this class of algorithms, the method of partial Gaussian elimination is used to reduce the original SDP into a smaller SDP instance. The smaller SDP instance is then solved using different approaches:

- *Wagner's approach*: The main idea of Wagner's approach [13], which has been applied to the syndrome decoding problem in [7], on a levels, is to partition the error vector into 2^a sub-vectors and to store them in a list together with their corresponding partial syndromes. After this, on each level, the lists are merged until a list of the solutions to the smaller SDP is obtained. Lastly, we go through this list to look for the solution to the original SDP.
- *Representations technique approach*: This technique was introduced in [10, 3], where we allow the sub-vectors of the error vector to overlap; this is called the subset sum representation technique. Due to the overlaps, the technique to merge two lists changes in this case.
- *Mixed approach (BJMM)*: In this approach, we allow both the above stated techniques. On the base level, we use the Wagner's approach and on the other levels we use the representation technique. This method for the Hamming metric was proposed in [3], also known as BJMM algorithm.

Following is the list of all the Lee metric ISD algorithms proposed so far:

- Lee-Brickell's algorithm and Stern's algorithm were adapted for codes over $\mathbb{Z}/4\mathbb{Z}$ in [8].
- Stern's algorithm was generalized for codes over $\mathbb{Z}/p^s\mathbb{Z}$ in [14], named as two-block algorithm.
- Furthermore, in [14], ISD algorithms based on new approaches, such as Wagner's algorithm, BJMM algorithm, and representation technique based algorithms, were adapted for codes over $\mathbb{Z}/p^s\mathbb{Z}$.
- In [5], the classical and quantum ISD algorithms based on Wagner's approach were presented for codes over finite fields, thus restricting to $\mathbb{Z}/p\mathbb{Z}$.
- s-blocks algorithm, based on the structure of the ring $\mathbb{Z}/p^s\mathbb{Z}$, was proposed in [14].
- Another algorithm based on the ring structure of $\mathbb{Z}/p^s\mathbb{Z}$ was proposed in [9], which splits the decoding problem over $\mathbb{Z}/p^s\mathbb{Z}$ into two successive decoding problems: the first one over $\mathbb{Z}/p^j\mathbb{Z}$ and the second one over $\mathbb{Z}/p^{s-j}\mathbb{Z}$ for some fixed $j \in \{1, 2, \ldots, s-1\}$.

Other techniques In the following, we present some ideas that may be used to either construct new decoding algorithms or speed-up ISD algorithms.

- *Scalar multiplication*: Recently in [2], the effect of scalar multiplication was studied for a random vector of a constant Lee weight. The results suggest that, using a suitable scalar $a \in \mathbb{Z}/p^s\mathbb{Z}$, the decoding of syndrome $a\mathbf{s}$ can be faster than the original $\mathbf{s}$ itself. This idea is specific for Lee metric, as in the Hamming metric case, scalar multiplication is a weight preserving map.
- *Lattice-based techniques*: Coding theory has strong resemblance with the lattice theory. In particular, the Lee metric can be seen as the metric induced by ℓ_1-norm. Due to this resemblance, new decoding techniques can be developed by adapting the lattice-reduction based algorithms from the ℓ_2-norm to the Lee metric.

Complexity In order to decode a code of length n over $\mathbb{Z}/q\mathbb{Z}$, the average-case time complexity of an ISD algorithm is given by

$$q^{(e(R,q)+o(1))n},$$

where R is the rate of the code, and the exponent $e(R,q)$ depends on the ISD algorithm. In Figure 1, we compare the asymptotic complexity of some ISD algorithms at different rates R by optimizing the internal parameters of each algorithm. From the comparison, we observe that the complexity of BJMM algorithm at level 2 is significantly lower than the other algorithms.

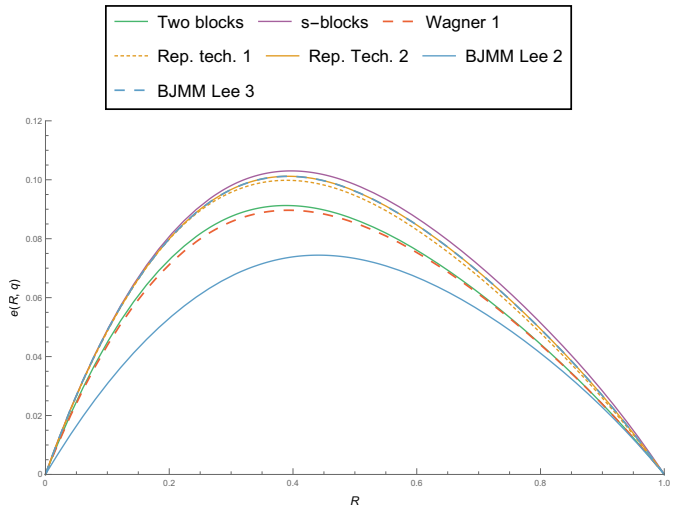


Fig. 1. Comparison of the asymptotic complexity of different Lee metric ISD algorithms for half-distance decoding of a free code over $\mathbb{Z}/7^2\mathbb{Z}$ that achieves the Gilbert-Varshamov bound.

ACKNOWLEDGEMENTS

The second author is supported by the Estonian Research Council grant number PRG49. The third author is supported by the Swiss National Science Foundation grant number 195290.

REFERENCES

[1] S. Barg. Some new NP-complete coding problems. *Problemy Peredachi Informatsii*, 30(3):23–28, 1994.

[2] Jessica Bariffi, Hannes Bartz, Gianluigi Liva, and Joachim Rosenthal. On the properties of error patterns in the constant lee weight channel. *arXiv preprint arXiv:2110.01878*, 2021.

[3] A. Becker, A. Joux, A. May, and A. Meurer. Decoding random binary linear codes in $2^{n/20}$: How $1+1=0$ improves information set decoding. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 520–536. Springer, 2012.

[4] E. Berlekamp, R. McEliece, and H. van Tilborg. On the inherent intractability of certain coding problems. *IEEE Trans. on Inf. Theory*, 24(3):384–386, May 1978.

[5] A. Chailloux, T. Debris-Alazard, and S. Etinski. Classical and quantum algorithms for generic syndrome decoding problems and applications to the lee metric. In J. H. Cheon and Jean-Pierre Tillich, editors, *Post-Quantum Cryptography*, pages 44–62, Cham, 2021. Springer International Publishing.

[6] L. Chen, Y.-K. Liu, S. Jordan, D. Moody, R. Peralta, R. Perlner, and D. Smith-Tone. Report on post-quantum cryptography. Technical Report NISTIR 8105, National Institute of Standards and Technology, 2016.

[7] M. Finiasz and N. Sendrier. Security bounds for the design of code-based cryptosystems. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 88–105. Springer, 2009.

[8] A.-L. Horlemann-Trautmann and V. Weger. Information set decoding in the Lee metric with applications to cryptography. *Advances in Mathematics of Communications*, online, 2019.

[9] T. S. C. Lau and C. H. Tan. On the design and security of lee metric mceliece cryptosystems. *Designs, Codes and Cryptography*, pages 1–23, 2022.

[10] A. May, A. Meurer, and E. Thomae. Decoding random linear codes in $\tilde{\mathcal{O}}(2^{0.054n})$. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 107–124. Springer, 2011.

[11] R. McEliece. A public-key cryptosystem based on algebraic coding theory. *DSN Progress Report*, pages 114–116, 1978.

[12] J. Stern. A method for finding codewords of small weight. In *International Colloquium on Coding Theory and Applications*, pages 106–113. Springer, 1988.

[13] D. Wagner. A generalized birthday problem. In *Annual International Cryptology Conference*, pages 288–304. Springer, 2002.

[14] V. Weger, K. Khathuria, A.-L. Horlemann-Trautmann, M. Battaglioni, P. Santini, and E. Persichetti. On the hardness of the Lee syndrome decoding problem. *arXiv preprint arXiv:2002.12785*, 2020.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On Cameron-Liebler sets of k-spaces in finite projective spaces (Part II)

Jonathan Mannaert *

* *Vrije Universiteit Brussel, Pleinlaan 2, B–1050 Brussel (e-mail: Jonathan.Mannaert@vub.be).*

Abstract: Cameron-Liebler sets of lines in a finite 3-dimensional space $\mathrm{PG}(3,q)$ originate from the study by Cameron and Liebler in 1982 of groups of collineations with equally many orbits on the points and the lines of $\mathrm{PG}(3,q)$. These objects have some interesting equivalent characterizations, and are examples of Boolean functions of degree one. In this talk, we focus on these objects and their generalisation from a geometric perspective, and report on several existence and non-existence results, including a lower bound on the existence of the parameter x (besides trivial examples).

Keywords: low degree Boolean functions, irreducible groups, Cameron-Liebler sets, extremal sets, strongly regular graphs.

1. INTRODUCTION

In Cameron and Liebler (1982) tried to classify specific orbits of subgroups of $\mathrm{PGL}(4,q)$ that admit the same number of orbits on lines then on points. In this study it was observed that these orbits on the lines satisfy diverse conditions. These line classes later obtained the name of *Cameron-Liebler line classes* in $\mathrm{PG}(3,q)$. These objects are in fact broader concepts then the original obits, yet they are nevertheless usefull in the classification of the subgroups. Each Cameron-Liebler line class $\mathcal{L}$ has a certain parameter x, for which holds that $|\mathcal{L}| = x(q^2+q+1)$. This integer x also denotes the intersection size of $\mathcal{L}$ with line spreads in $\mathrm{PG}(3,q)$.
Since a classification of these orbits or line classes would aid their research question, this classification was briefly conducted in Cameron and Liebler (1982). Here they classified all Cameron-Liebler line classes of parameter $x \in \{0,1,2,q^2-1,q^2,q^2+1\}$. They also conjectured that the only line classes of this form that exist are: (1) the empty set, (2) all lines through a point, (3) all lines in a plane, (4) all lines through in a plane π or all lines through a point $p \notin \pi$ and (4) the complements of the previous examples. These examples are also known as the *trivial examples*. The conjecture was later disproven in Drudge (1999) who gave a first example of a non-trivial Cameron-Liebler line class in $\mathrm{PG}(3,3)$ of parameter $x = 5$. This example was later generalized to an infinite famely in $\mathrm{PG}(n,q)$ for n odd in A. Bruen and Drudge (1999)

Now we switch the viewpoint to coding theory. Let $n \geq 1$ and A be a set of q symbols. The Hamming graph $H(n,q)$ is the graph with vertex set the set of words of length n over A and two vertices being adjacent if and only if their Hamming distance is 1. The Hamming graph $H(n,2)$ is the hypercube. Clearly, a q-ary code of length n can be considered as a subset of the vertex set of $H(n,q)$. So it is quite natural to translate properties of the code C into graph theoretical properties. Conversely, it is natural to define codes as substructures in graphs different from the Hamming graph as well, replacing the Hamming distance by the graph distance. Let C be a code in a regular graph Γ with vertex set V. We follow the definition found in e.g. Neumaier (1992). Let x be any vertex of Γ, then $d(x,C) = \min\{d(x,y)|y \in C\}$. The covering radius $\rho = \max\{d(x,C)|x \in \Gamma\}$, it is the minimal integer ρ such that the spheres of radius ρ around the codewords of C cover the vertices of Γ. For a code C, minimum distance $d(C)$ and covering radius $\rho(C)$ are related by $d(C) \leq 2\rho(C)+1$. The code C is called *perfect* in case of equality, which is equivalent with the property that the spheres with radius $\rho(C)$ around the codewords partition the vertex set V.

Completely regular codes have been introduced in Delsarte (1973) as a generalization of perfect codes. Assume that C is a code in a distance regular graph. Let $C_i = \{x \in \Gamma | d(x,C) = i\}$, then $C_i \neq \emptyset \iff 0 \leq i \leq \rho(C)$, and $C_0 = C$. The sets C_i partition the vertex set of Γ. The code C is called *completely regular* if every vertex $x \in C_i$ has a constant number of neighbors a_i, b_i, respectively c_i in C_{i-1}, C_i, respectively C_{i+1}. This is equivalent with the partition of the vertex set of Γ into the components C_i being equitable.

In I.Mogilnykh (2022) a very nice connection is made between complete regular codes and Cameron-Liebler line classes. In order to understand this connection, we require the definition of a t-desing. In general a t-(v,k,λ) desing, or t-design for short, is a set of points and blocks with v points, each block contains k points and through every t points there are λ blocks. It is discussed that Cameron-Liebler line classes $\mathcal{L}$ in $\mathrm{PG}(n,q)$ are in fact special completly regular codes where the points of $\mathrm{PG}(n,q)$ and the lines of $\mathcal{L}$ do not admit to the blocks of a t-desing. It can be shown that these codes have covering radius 1.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2. A GENERALISATION TO N-DIMENSIONAL PROJECTIVE SPACES

Let $q = p^h$ be a prime power, and denote the finite field of order q as $\mathrm{GF}(q)$. The n-dimensional projective space over $\mathrm{GF}(q)$ is the geometry consisting of all i-dimensional subspaces of the $n+1$-dimensional vector space $V(n+1,q)$ over $\mathrm{GF}(q)$, $1 \leq i \leq d$, and is denoted by $\mathrm{PG}(n,q)$. The fundamental theorem of projective geometry states that all isomorphisms of $\mathrm{PG}(n,q)$ are induced by the semi-linear maps of the underlying vector space. Note that we call the i-dimensional subspaces of $V(n+1,q)$, $1 \leq i \leq n$, the points, lines, planes, etc. of $\mathrm{PG}(n,q)$. In general we define the set of i-spaces in $\mathrm{PG}(n,q)$ by Π_i.
The concepts of Cameron-Liebler line classes have been generalized to $\mathrm{PG}(n,q)$, see Blokhuis et al. (2018) and Rodgers et al. (2018). In order to do so, we will need the concept of a *characteristic vector* of a set of k-spaces $\mathcal{S}$. This $0,1$ valued vector of size $|\Pi_k|$ has a one on a certain position if and only if the corrsponding k-spaces lies inside $\mathcal{S}$. Using this vector, we can define Cameron-Liebler sets of k-spaces.

2.1 Definition

Suppose that P_n is the point-(k-space) incidence matrix of $\mathrm{PG}(n,q)$. This matrix is the $0,1$ values matrix where the rows correspond with the points and the columns with the k- dimensional subspaces of $\mathrm{PG}(n,q)$. A certain position (p,K) has value one if and only if $p \in K$. Using this matrix, a *Cameron-Liebler set of k-spaces* in $\mathrm{PG}(n,q)$ is a set of k-spaces such that the characteristic vector is contained in $\mathrm{Im}(P_n^T)$.
We say that the Cameron-Liebler set of k-spaces $\mathcal{L}$ has parameter x, if and only if

$$|\mathcal{L}| = x \begin{bmatrix} n \\ k \end{bmatrix}_q .$$

Recall the Gaussian binomial coefficient, for $a \leq b$ be natural numbers,

$$\begin{bmatrix} b \\ a \end{bmatrix}_q = \frac{(q^b-1)\dots(q^{b-a+1}-1)}{(q^a-1)\dots(q-1)}$$

which represents the number of $(a-1)$-dimensional projective spaces in a projective space of dimension $b-1$. This definition automatically implies that $0 \leq x \leq \frac{q^{n+1}-1}{q^{k+1}-1}$.

2.2 Connection with Boolean functions

Another way to formalize the definition of these Cameron-Liebler sets of k-spaces is by using *Boolean degree 1 functions*. A Boolean degree 1 function f on the k-spaces of $\mathrm{PG}(n,q)$, is a boolean function that is a linear combination of boolean functions of point-pencils, i.e. sets of k-spaces through a fixed point. In particular, this means that

$$f = \sum_{p \in \mathrm{PG}(n,q)} c_p x_p^+,$$

Here c_p are constants and x_p^+ corresponds to the boolean function of the set $[p]_k := \{K \in \Pi_k \mid p \in K\}$. Consequently, we have that if χ_p is the characteristic vector of $[p]_k$ then $x_p^+(K) = \chi_p(K)$. Cameron-Liebler sets of k-spaces and Boolean degree 1 functions are in fact equivalent statements. This implies that every result of one can be transformed to the the other. Yet studying these objects in both ways can yield powerfull results. For more information we refer to Filmus and Ihringer (2019).

2.3 Other properties and examples

Besides this definition other equivalent versions are also known. The following Theorem will list some of those, here a *k-spread* $\mathcal{S}$ denotes a set of k-spaces that partition the point in $\mathrm{PG}(n,q)$. In particular, it is know that these spreads exist if and only if $(k+1) \mid (n+1)$.

Theorem 1. (Blokhuis et al., 2018, Theorem 2.2) Let $\mathcal{L}$ be a non-empty set of k-spaces in $\mathrm{PG}(n,q)$, $n \geq 2k+1$, with characteristic vector $\chi_{\mathcal{L}}$, and x so that $|\mathcal{L}| = x \begin{bmatrix} n \\ k \end{bmatrix}_q$. Then the following properties are equivalent.

(1) $\chi_{\mathcal{L}} \in \mathrm{Im}(P_n^T) = (\ker(P_n))^{\perp}$, with P_n the point-(k-space) incidence matrix of $\mathrm{PG}(n,q)$.
(2) For every k-space K, the number of elements of $\mathcal{L}$ disjoint from K is equal to $(x - \chi_{\mathcal{L}}(K)) \begin{bmatrix} n-k-1 \\ k \end{bmatrix}_q q^{k^2+k}$.
(3) If $k+1 \mid n+1$, i.e. if and only if $\mathrm{PG}(n,q)$ has k-spreads, then $|\mathcal{L} \cap \mathcal{S}| = x$ for any k-spread $\mathcal{S}$.

It is also clear that this theorem induces that if $(k+1) \mid (n+1)$, then x is in fact a positive integer. This condition only holds for this particular case. Because as we will see in the next example, for $(k+1) \nmid (n+1)$ we always find non-integer parameters.

Fact 2. The *trivial* examples of Cameron-Liebler line classes in $\mathrm{PG}(3,q)$ are easily generalized to sets of k-spaces in $\mathrm{PG}(n,q)$. In particular, using Theorem 1, we can obtain the following examples.

(1) The empty set is a Cameron-Liebler set of k-spaces of parameter $x = 0$.
(2) A point-pencil is an example of parameter $x = 1$.
(3) The set of k-spaces inside a hyperplane is an example of parameter $x = \frac{q^{n-k}-1}{q^{k+1}-1}$.
(4) The disjoint union of examples (2) and (3) is a Cameron-Liebler set of k-spaces of parameter $x = \frac{q^{n-k}-1}{q^{k+1}-1} + 1$.
(5) Complements of these examples are also Cameron-Liebler sets of k-spaces of parameter $\frac{q^{n+1}-1}{q^{k+1}-1} - x$.

As previously noted, it can be seen that examples (3), (4) and (5) have a parameter that is only an integer if and only if $(k+1) \mid (n+1)$.

3. NON-EXISTENCE RESULTS

In this section we consider some known non-existence results and classifications. We combine the results that were obtained for Boolean degree 1 functions and Cameron-Liebler sets directly. The following theorem follows (partially) immediately from Theorem 1.

Theorem 3. (Blokhuis et al., 2018, Theorem 4.3) There do not exist Cameron-Liebler k-sets in $\mathrm{PG}(n,q)$ of parameter $x \in]0,1[$ and if $n \geq 3k+2$, then there are no Cameron-Liebler sets of k spaces with parameter $x \in]1,2[$.

In order to prove that $x \notin]0,1[$, we can simply consider Theorem 1 (2) and obtain that for these values the number of skew k-spaces of $\mathcal{L}$ to a fixed k-space K is a negative number. This is a contradiction. A second classification results makes use from some Erdös-Ko-Rado results and is as follows.

Theorem 4. (Blokhuis et al., 2018, Theorem 4.1) Let $\mathcal{L}$ be a Cameron-Liebler set of k-spaces with parameter $x = 1$ in $\mathrm{PG}(n,q)$, $n \geq 2k+1$. Then $\mathcal{L}$ consists out of all the k-spaces through a fixed point or $n = 2k+1$ and $\mathcal{L}$ is the set of all the k-spaces in a hyperplane of $\mathrm{PG}(2k+1,q)$.

Besides these rather small classifications, there also exist stronger non-existence conditions. These are in fact the results which we want to improve.

Theorem 5. (Blokhuis et al., 2018, Theorem 4.9) There are no Cameron-Liebler sets of k-spaces in $\mathrm{PG}(n,q)$, with $n \geq 3k+2$ (and $q \geq 3$), of parameter x if

$$2 \leq x \leq \frac{1}{\sqrt[8]{2}} q^{\frac{n}{2}-\frac{k^2}{4}-\frac{3k}{4}-\frac{3}{2}} (q-1)^{\frac{k^2}{4}-\frac{k}{4}+\frac{1}{2}} \sqrt{q^2+q+1}.$$

Note that this upper bound has size roughly $q^{\frac{n}{2}-k}$.

Fact 6. The condition that $q \geq 3$ is in brackets because this was in the original theorem, but it has been proven in Filmus and Ihringer (2019) that for $q \leq 3$ all Cameron-Liebler sets of k-spaces are trivial examples.

Theorem 7. (Ihringer, 2020, Theorem 7) Let $n \geq 2k+1$ and suppose that $\mathcal{L}$ is a Cameron-Liebler set of k-spaces in $\mathrm{PG}(n,q)$ of parameter x. If $16x \leq \min\{q^{\frac{n-k-l+2}{3}}, q^{\frac{n-2k-r}{3}}\}$, where $n+1 = m(k+1)-r$ with $0 \leq r < k+1$ and $\frac{q^{l-1}-1}{q-1} < x \leq \frac{q^l-1}{q-1}$, then $x \leq 2$ and $\mathcal{L}$ is trivial.

This bound would in the best case scenario have size roughly $\min\{q^{\frac{n-k+1}{3}}, q^{\frac{n-2k}{3}}\} = q^{\frac{n-2k}{3}}$.
In general no theorem of these above excludes the other, because both have different strengths in different cases of (n,k).

4. NEW RESULTS

This is joint wok with Jan De Beule and Leo Storme. In order to soften the notation, we will denote $[\pi]_k$ as the set of k-spaces in $\mathrm{PG}(n,q)$ that are contained in the i-dimensional subspace π, $i > 0$. Similarly, $[p]_p$ denotes the set of k-spaces containing the point p.

4.1 Induces Cameron-Liebler sets

The main idea for this improved lower bound on x comes from Boolean degree 1 functions, see Filmus and Ihringer (2019), and the thesis of Drudge, see Drudge (1998). Both describe a version of the following theorem. In Filmus and Ihringer (2019) it is described more subtly, while in Drudge (1998) it is only denoted for lines.

Theorem 8. (Folklore). Consider a Cameron-Liebler sets of k-spaces $\mathcal{L}$ in $\mathrm{PG}(n,q)$, with $n \geq 2k+1$, and let π be a subspace of dimension $i \geq k+1$. Then $\mathcal{L} \cap [\pi]_k$ is also a Cameron-Liebler set of k-spaces in π.

Proof. We will show this using Boolean functions. Suppose that $\chi_{\mathcal{L}}$ is a Boolean degree 1 function, then we have, with a similar notation as above, that

$$\chi_{\mathcal{L}} = \sum_{p \in \mathrm{PG}(n,q)\setminus\pi} c_p x_p^+ + \sum_{p \in \pi} c_p x_p^+.$$

The restriction of $\chi_{\mathcal{L}}$ to π is cleary $\sum_{p\in\pi} c_p x_p^+$.

The main idea, using this theorem, is to consider the Cameron-Liebler sets of k-spaces that are induced in subspaces and try to translate classification results using this connection. However, we should be very carefull because Theorem 8 does not give a connection between the parameter of $\mathcal{L}$ and the parameter of the induced set. This will be our focus now.

4.2 Connecting the parameter with parameters of induced sets

The main result of this section is the following lemma.

Lemma 9. (De Beule et al., 2022, Lemma 4.1) Suppose that $\mathcal{L}$ is a Cameron-Liebler set of k-spaces in $\mathrm{PG}(n,q)$ of parameter x. Then for every t, such that $2k+1 \leq t \leq n-1$ and $n \geq 2k+2$, and an arbitrary k-space K in $\mathrm{PG}(n,q)$,

$$x = \frac{\left(\sum_{K\in\pi_i} x_{\pi_i} - \begin{bmatrix} n-k \\ t-k \end{bmatrix}_q \chi_{\mathcal{L}}(K)\right)}{\begin{bmatrix} n-k-1 \\ t-k-1 \end{bmatrix}_q} + \chi_{\mathcal{L}}(K), \qquad (1)$$

where $\chi_{\mathcal{L}}(K)$ is the value of the characteristic vector of $\mathcal{L}$ at position K, x_{π_i} the parameter of the Cameron-Liebler k-set induced in π_i, and where the sum runs over all t-spaces π_i through K.

Proof. The proof is based on a counting argument, where we count the pairs (K', π), with $K' \in \mathcal{L}$ a k-dimensional subspace inside π skew to K, and π a t-dimensional subspace containing both K and K'. Important to know is that the number of skew k-spaces to K is denoted in Theorem 1 (2) and since $\mathcal{L}$ restricted to every t-space π also induces a Cameron-Liebler set of k-space of a certain parameter x_π a similar number of skew k-spaces can be found depending on x_π.

4.3 Using Classification results in smaller subspaces

The strength of Lemma 9 is the fact that t can be chosen freely within the possible range. This implies that we can use some non-existence conditions while playing with this dimension t. In general we have that if we choose $K \in \mathcal{L} \neq \emptyset$ that all $x_{\pi_i} \neq 0$, since all induced sets are not empty. In addition, we also have, due to Theorem 3 that $x_{\pi_i} \geq 1$. Now we will make use of the following theorem, which is a generalisation of a result from Drudge Drudge (1998) who proved this for lines.

Theorem 10. (De Beule et al., 2022, Theorem 3.4) Let $n \geq 2k+1$. Suppose that $\mathcal{L}$ is a Cameron-Liebler set of k-spaces in $\mathrm{PG}(n,q)$, such that there exists an i-space π, with $i \geq k+1$, and such that $\mathcal{L} \cap [\pi]_k$ consists of the set of k-spaces through a point $p \in \pi$. Then $\mathcal{L}$ is the set of k-spaces through this same point p.

An important consequence of this theorem is the following. Let $t > 2k+1$ and suppose that there exist a $x_{\pi_i} = 1$, then from Theorem 4 we obtain that the induced Cameron-Liebler set is a point-pencil. Consequently, using Theorem 10 we have that $\mathcal{L}$ is a point-pencil. So if we pose the restriction that $\mathcal{L}$ is not a point-pencil, we can say that

every $x_{\pi_i} \neq 1$.
Finaly, if $t \geq 3k+2$, we also obtain, by Theorem 3 that $x_{\pi_i} \geq 2$.
Filling in these bounds in Lemma 9, we obtain that

$$x \geq \frac{\begin{bmatrix} n-k \\ t-k \end{bmatrix}_q}{\begin{bmatrix} n-k-1 \\ t-k-1 \end{bmatrix}_q} + 1 = \frac{\begin{bmatrix} n-k \\ 2k+2 \end{bmatrix}_q}{\begin{bmatrix} n-k-1 \\ 2k+1 \end{bmatrix}_q} + 1 = \frac{q^{n-k}-1}{q^{2k+2}-1} + 1,$$

where in the last two formulas, we let $t = 3k+2$.

Fact 11. The reason why $t = 3k+2$ is due to the fact that this value gives the best lower bound. Filling in $t > 3k+2$ would only increase the nominator of the previous equation.

This results in the following theorem.

Theorem 12. (De Beule et al., 2022, Theorem 6.2) Suppose that $n \geq 3k+3$ and $k \geq 1$. Let $\mathcal{L}$ be a Cameron-Liebler set of k-spaces of parameter x in $\mathrm{PG}(n,q)$ such that $\mathcal{L}$ is not a point-pencil, nor the empty set. Then it holds that

$$x \geq \frac{q^{n-k}-1}{q^{2k+2}-1} + 1.$$

4.4 A direct improvement

Theorem 12 was stated in De Beule et al. (2022) but can be improved using Theorem 5. Due to Fact 6, this follows for every q and $t \geq 3k+2$. Filling in this better bound for every x_{π_i}, we obtain that

$$x_{\pi_i} \geq C(t,k,q),$$

with $C(t,k,q) = \frac{1}{\sqrt[8]{2}} q^{\frac{t}{2} - \frac{k^2}{4} - \frac{3k}{4} - \frac{3}{2}} (q-1)^{\frac{k^2}{4} - \frac{k}{4} + \frac{1}{2}} \sqrt{q^2+q+1}$.
Again we choose $t = 3k+2$. This results in the following Theorem.

Theorem 13. Suppose that $n \geq 3k+3$ and $k \geq 1$. Let $\mathcal{L}$ be a Cameron-Liebler set of k-spaces of parameter x in $\mathrm{PG}(n,q)$ such that $\mathcal{L}$ is not a point-pencil, nor the empty set. Then it holds that

$$x \geq C(k,q) \left(\frac{q^{n-k}-1}{q^{2k+2}-1} \right) + 1,$$

for $C(k,q) = \frac{1}{\sqrt[8]{2}} q^{\frac{3k}{4} - \frac{k^2}{4} - \frac{1}{2}} (q-1)^{\frac{k^2}{4} - \frac{k}{4} + \frac{1}{2}} \sqrt{q^2+q+1} - 1$.

This lower bound is roughly of size $q^{n - \frac{3k}{2} - \frac{1}{2}}$, which is a direct improvement of Theorem 5 and Theorem 7 if $n \geq 3k+2$, so for all possible values of (n,k).

REFERENCES

A. Bruen, A. and Drudge, K. (1999). The construction of Cameron–Liebler line classes in $\mathrm{PG}(3,q)$. *Finite Fields Appl*, 5(1), 35–45. doi:10.1006/ffta.1998.0239. URL `https://doi.org/10.1006/ffta.1998.0239`.

Blokhuis, A., De Boeck, M., and D'haeseleer, J. (2018). Cameron-Liebler sets of k-spaces in $\mathrm{PG}(n,q)$. *Des. Codes Cryptogr.* URL `https://doi.org/10.1007/s10623-018-0583-1`.

Cameron, P. and Liebler, R. (1982). Tactical decompositions and orbits of projective groups. *Linear Algebra Appl.*, 46, 91–102. doi:10.1016/0024-3795(82)90029-5. URL `https://doi.org/10.1016/0024-3795(82)90029-5`.

De Beule, J., Mannaert, J., and Storme, L. (2022). Cameron–liebler k-sets in subspaces and non-existence conditions. *Des. Codes and Cryptogr.* doi: https://doi.org/10.1007/s10623-021-00995-0.

Delsarte, P. (1973). An algebraic approach to the association schemes of coding theory. *Philips Res. Rep. Suppl.*, (10), vi+97.

Drudge, K. (1998). *Extremal sets in projective and polar spaces*. Ph.D. thesis, The University of Western Ontario, London, Canada.

Drudge, K. (1999). On a conjecture of Cameron and Liebler. *European J. Combin.*, 20(4), 263–269. doi:10.1006/eujc.1998.0265. URL `https://doi.org/10.1006/eujc.1998.0265`.

Filmus, Y. and Ihringer, F. (2019). Boolean degree 1 functions on some classical association schemes. *J. Comb. Theory, Ser. A*, 162, 241–270.

Ihringer, F. (2020). Remarks on the Erdős Matching Conjecture for Vector Spaces. *arXiv e-prints*, arXiv:2002.06601.

I.Mogilnykh (2022). Completely regular codes in johnson and grassmann graphs with small covering radii. *Elec. J. Combin.*, 29 (2), P2.57.

Neumaier, A. (1992). Completely regular codes. *Discrete Math.*, 106–107, 353–360. doi:10.1016/0012-365X(92)90565-W. URL `https://doi.org/10.1016/0012-365X(92)90565-W`.

Rodgers, M., Storme, L., and Vansweevelt, A. (2018). Cameron-Liebler k-classes in $\mathrm{PG}(2k+1,q)$. *Combinatorica*, 38(3), 739–757. doi:10.1007/s00493-016-3482-y. URL `https://doi.org/10.1007/s00493-016-3482-y`.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Positive semidefinite quadratic forms on varieties defined by quadratic forms

Sarah Hess *

* *University of Konstanz, 78457 Konstanz, Germany*
(e-mail: sarah.hess@uni-konstanz.de).

Abstract: For a fixed number of $n+1$ ($n \geq 1$) variables and even degree $2d$ ($d \geq 1$), the SOS cone $\Sigma_{n+1,2d}$ of all real forms representable as finite sums of squares (SOS) of half degree d real forms is included in the PSD cone of all positive semidefinite (PSD) real forms $\mathcal{P}_{n+1,2d}$. Hilbert (1888) states that both cones coincide if and only if $n+1=2$, $d=1$ or $(n+1,2d)=(3,4)$. In this talk, we discuss necessary or sufficient conditions to extend local positive semidefiniteness of real quadratic forms along projective varieties generated by s ($s \geq 0$) real quadratic forms. Those conditions allow us to construct an explicit filtration of intermediate cones $\Sigma_{n+1,2d} = C_0 \subseteq C_1 \subseteq \ldots \subseteq C_{s-1} \subseteq C_s = \mathcal{P}_{n+1,2d}$ (between the SOS and PSD cone) along the Veronese variety. Indeed, the latter is known to be a projective variety finitely induced by real quadratic forms. We analyze this filtration for proper inclusions. In fact, after applying an inductive argument, it suffices to investigate the situation for a truncated subfiltration of the former. A result of Blekherman et al. (2016) on projective varieties of minimal degree permits us to handle the inclusion $C_0 \subseteq C_1$. Generalizing this observation, we are able to show $\Sigma_{n+1,2d} = C_0 = \ldots = C_n$. Finally, we lay out the situation in the basic non Hilbert case of quaternary quartics by identify exactly two strictly separating intermediate cones in the particular filtration of $\Sigma_{4,4}$ and $\mathcal{P}_{4,4}$ via considerations of real forms based on techniques due to Robinson (1969) and Choi and Lam (1977a,b). This is a work in progress with Salma Kuhlmann and Charu Goel.

Keywords: Real quadratic forms, projective varieties generated by real quadratic forms, positive semidefinite forms, sums of squares, intermediate cones, varieties of minimal degree

1. INTRODUCTION

For $n \geq 0$, let $\mathbb{R}[X]$ be the polynomial ring in $n+1$ variables with coefficients in $\mathbb{R}$. If all monomials appearing in $f \in \mathbb{R}[X]$ are of the same total degree d ($d \geq 1$), then f is a *(real) form (of total degree d)*. The set of all real forms of total degree d in $\mathbb{R}[X]$ is $\mathcal{F}_{n+1,d}$. In particular, $f \in \mathcal{F}_{n+1,2}$ is a *(real) quadratic form*. Moreover, if for $f \in \mathcal{F}_{n+1,2d}$ there exist some $t \geq 1$ and $g_1, \ldots, g_t \in \mathcal{F}_{n+1,d}$ such that $f = \sum_{i=1}^{t} g_i^2$, then f is a *sum of squares (SOS)*. The cone of all SOS forms in $\mathcal{F}_{n+1,2d}$ is $\Sigma_{n+1,2d}$. Moreover, $f \in \mathcal{F}_{n+1,2d}$ is *locally positive semidefinite on* $W \subseteq \mathbb{R}^{n+1}$ if $f(x) \geq 0$ holds for all $x \in W$. In this case we write $f|_W \geq 0$, respectively, $f \geq 0$ for $W = \mathbb{R}^{n+1}$. In the latter case, f is *(globally) positive semidefinite (PSD)*. The cone of all PSD forms in $\mathcal{F}_{n+1,2d}$ is $\mathcal{P}_{n+1,2d}$. It is clear that $\Sigma_{n+1,2d} \subseteq \mathcal{P}_{n+1,2d}$ always holds true and, especially, $\Sigma_{1,2d} = \mathcal{P}_{1,2d}$ in the univariate case (see Marshall (2008)). However, the situation is more evolved in the multivariate cases. Hence, from now on we assume $n \geq 1$.

Theorem 1. (Hilbert (1888)) Let n and d be positive integers. Then $\Sigma_{n+1,2d} = \mathcal{P}_{n+1,2d}$ if and only if $n+1=2$ or $d=1$ or $(n+1,2d)=(3,4)$.

All cases in which the SOS and PSD cone coincide are called *Hilbert cases*, whereas all others are refered to as *non Hilbert cases*. The two simplest non Hilbert cases $(3,6)$ and $(4,4)$ are the *basic* non Hilbert cases.

Let $(n+1,2d)$ from now on denote a non Hilbert case and $\{m_0(X), \ldots, m_k(X)\}$ be an ordered monomial basis of $\mathcal{F}_{n+1,d}$ with $k := \dim(\mathcal{F}_{n+1,d}) - 1$. For $l \in \{n,k\}$, let $\mathbb{P}^l$ be the l-dimensional projective space of the complex numbers and the set of all real points of W is denoted by $W(\mathbb{R})$ for any $W \subseteq \mathbb{P}^l$. Implicitly $x \in \mathbb{R}^{l+1}$ is assumed for any $[x] \in W(\mathbb{R})$. A form $f \in \mathcal{F}_{l+1,2d}$ is *locally positive semidefinite on* $W(\mathbb{R}) \subseteq \mathbb{P}^n(\mathbb{R})$ if $f(x_0, \ldots, x_l) \geq 0$ holds for any $[x_0 : \ldots : x_l] \in W(\mathbb{R})$ and we write $f|_{W(\mathbb{R}))} \geq 0$. This is a well defined expression due to the homogeneity of f in even degree. In particular, the cone of all forms in $\mathcal{F}_{l+1,2d}$ which are locally positive semidefinite on $\mathbb{P}^n(\mathbb{R})$ is the former PSD cone $\mathcal{P}_{l+1,2d}$.
In a Gram matrix approach (see Choi et al. (1995), Powers and Wörmann (1998)), we consider the isomorphism

$$Q\colon \mathrm{Sym}_{k+1}(\mathbb{R}) \to \mathcal{F}_{k+1,2}$$
$$A \mapsto q_A,$$

where $q_A(Z_0, \ldots, Z_k) := (Z_0 \ldots Z_k) A (Z_0 \ldots Z_k)^t$, and the surjective linear *Gram map*

$$\mathcal{G}\colon \mathrm{Sym}_{k+1}(\mathbb{R}) \to \mathcal{F}_{n+1,2d}$$
$$A \mapsto f_A,$$

where $f_A(X) := (m_0(X)\dots m_k(X))A(m_0(X)\dots m_k(X))^t$ for the indeterminantes $X = (X_0, \dots, X_n)$, over the $\mathbb{R}$-vector space $\mathrm{Sym}_{k+1}(\mathbb{R})$ of real symmetric $(k+1)\times(k+1)$ matrices. Then a generic $A_f \in \mathcal{G}^{-1}(f)$ for any $f \in \mathcal{F}_{n+1,2d}$ can be fixed. In fact, any $A \in \mathcal{G}^{-1}(f)$ is a *Gram matrix associated to* f and for any such, $q_A := Q(A) \in \mathcal{F}_{k+1,2}$ is a *(real) quadratic form associated to* f.

Proposition 2. A form $f \in \mathcal{F}_{n+1,2d}$ is SOS if and only if there exists a real quadratic form associated to f which is locally positive semidefinite on $\mathbb{P}^k(\mathbb{R})$.

Under the consideration of the *(projective) Veronese embedding*

$$\begin{aligned} V\colon \mathbb{P}^n &\to \mathbb{P}^k \\ [x] &\mapsto [m_0(x) : \dots : m_k(x)] \end{aligned}$$

and its image the *(projective) Veronese variety* $V(\mathbb{P}^n)$, the PSD forms in $\mathcal{F}_{n+1,2d}$ can be characterized.

Proposition 3. A form $f \in \mathcal{F}_{n+1,2d}$ is PSD if and only if there exists a real quadratic form associated to f which is locally positive semidefinite on $V(\mathbb{P}^n)(\mathbb{R})$.

2. THE MAIN QUESTIONS

The previous two propositions reveal that the question of whether or not a given PSD form is SOS is equivalent to asking whether or not a given locally on $V(\mathbb{P}^n)(\mathbb{R})$ positive semidefinite real quadratic form can be extended to a real quadratic form locally positive semidefinite on $\mathbb{P}^k(\mathbb{R})$ over the set of real points of the Veronese variety. Indeed, the latter is a projective variety finitely generated by real quadratic forms of a specific structure imposed by the Gram map (see Plaumann (2020)). More precisely, the projective variety $V(\mathbb{P}^n)$ is induced by

$$\begin{aligned} \mathcal{S} := \{q(Z_0, \dots, Z_k) := Z_iZ_j - Z_sZ_t \mid \mathrm{LE}(m_i) + \mathrm{LE}(m_j) \\ = \mathrm{LE}(m_s) + \mathrm{LE}(m_t)\} \subseteq \mathbb{R}[Z_0, \dots, Z_k], \end{aligned}$$

where LE denotes the (leading) exponent of the indicated monomial. In general, the following question has to be answered.

Question 1. Let $W_0 \subseteq W_1$ be projective varieties finitely induced by real quadratic forms with non-empty sets of real points. Assume that a real quadratic form q is locally positive semidefinite on $W_0(\mathbb{R})$. When exactly does there exist a real quadratic form q_0 vanishing on $W_0(\mathbb{R})$ such that $q + q_0$ is locally positive semidefinite on $W_1(\mathbb{R})$?

Under the assumption of W_0 being an irreducible projective variety with Zariski dense set of real points $W_0(\mathbb{R})$ and W_1 being the projective space $\mathbb{P}^k$, Blekherman et al. (2016) give an answer to the above question. They establish that any real quadratic form q which is locally positive semidefinite on $W_0(\mathbb{R})$ is already SOS in the respective real homogeneous coordinate ring if and only if W_0 is a projective variety of minimal degree, i.e. a nondegenerate (not contained in any hyperplane of $\mathbb{P}^k$) irreducible projective variety with $\deg(W_0) = 1 + \mathrm{codim}(W_0)$. This result provides an alternative proof of Hilbert's 1888 Theorem by setting W_0 to be the Veronese variety and observing it being a projective variety of minimal degree exactly in the Hilbert cases.

Any subset $\mathcal{S}'$ of $\mathcal{S}$ naturally induces a supvariety W of the Veronese variety, which is consequently again finitely induced by real quadratic forms of the specific structure imposed by the Gram map. The kernel of the Gram map can be described via the set of real points of the Veronese variety. Indeed, the set $Q(\mathcal{G}^{-1}(f))$ of all quadratic forms associated to $f \in \mathcal{F}_{n+1,2d}$ is completely determined by

$$\mathcal{G}^{-1}(f) = \{A \in \mathrm{Sym}_{k+1}(\mathbb{R}) \mid q_A = q_{A_f} \text{ on } V(\mathbb{P}^n)(\mathbb{R})\}.$$

Hence, given a quadratic form locally positive semidefinite on $W(\mathbb{R})$, we can ask under exactly what conditions this form extends to a quadratic form locally positive semidefinite on $\mathbb{P}^k(\mathbb{R})$ over the set of real points of the Veronese variety. Set

$$\begin{aligned} C_W &:= \{f \in \mathcal{F}_{n+1,2d} \mid \exists A \in \mathcal{G}^{-1}(f)\colon q_A|_{W(\mathbb{R})} \geq 0\} \\ &= \{f \in \mathcal{F}_{n+1,2d} \mid \exists A \in \mathcal{G}^{-1}(f)\colon q_A|_{W(\mathbb{R})} \geq 0 \\ &\qquad \wedge\ q_A = q_{A_f} \text{ on } V(\mathbb{P}^n)(\mathbb{R})\}. \end{aligned}$$

Then by Proposition 2 and Proposition 3, it is clear that C_W is an intermediate cone of the SOS and PSD cone. We especially investigate the inclusions in

$$\Sigma_{n+1,2d} \subseteq C_W \subseteq \mathcal{P}_{n+1,2d}$$

for strictness. Indeed, at least one of these inclusions has to be strict because $(n+1, 2d)$ is assumed to be a non Hilbert case. The following question has to be answered.

Question 2. Let $W_0 \subseteq W_1 \subseteq W_2$ be projective varieties finitely induced by real quadratic forms with non-empty sets of real points. Assume that a real quadratic form q is locally positive semidefinite on $W_1(\mathbb{R})$. When exactly does there exist a real quadratic form q_0 vanishing on $W_0(\mathbb{R})$ such that $q + q_0$ is locally positive semidefinite on $W_2(\mathbb{R})$?

3. A FILTRATION OF INTERMEDIATE CONES

We algorithmically construct a particular $\mathcal{S}' \subseteq \mathcal{S}$ with a fixed numeration $\mathcal{S}' = \{p_1, \dots, p_s\}$ ($s := \#\mathcal{S}'$) such that the zero set of $\mathcal{S}'$ is the Veronese Variety and

$$V(\mathbb{P}^n) = W_s \subsetneq W_{s-1} \subsetneq \dots \subsetneq W_1 \subsetneq W_0 = \mathbb{P}^k$$

for $W_i := \mathcal{V}(p_1, \dots, p_i)$ ($i \in \{1, \dots, s\}$) and $W_0 := \mathbb{P}^k$. This leads to a corresponding strict filtration of sets of real points

$$V(\mathbb{P}^n)(\mathbb{R}) = W_s(\mathbb{R}) \subsetneq \dots \subsetneq W_0(\mathbb{R}) = \mathbb{P}^k(\mathbb{R}).$$

Setting $C_i := C_{W_i}$ we thus obtain a filtration of intermediate cones of the SOS and PSD cone, namely

$$\Sigma_{n+1,2d} = C_0 \subseteq C_1 \subseteq \dots \subseteq C_{s-1} \subseteq C_s = \mathcal{P}_{n+1,2d} \quad (1)$$

(see Goel (2020)). Since $(n+1, 2d)$ is a non Hilbert case by choice, at least one inclusion in (1) has to be strict. An answer to Question 2 in particularly provides a tool for identifying all strict inclusions in (1). Yet, it is not compulsory to investigate each inclusion in (1).

In the explicit construction of $\mathcal{S}'$, we determine

$$s = \#\mathcal{S}' = \sum_{m=1}^{n} 2(k(m) - m) - (k(m-1) + 1)$$

with

$$\begin{aligned} k\colon \mathbb{Z}_{\geq 0} &\to \mathbb{Z}_{\geq 0} \\ m &\mapsto \binom{m+d}{m} - 1. \end{aligned}$$

Setting $\tau := 2(k(n) - n) - (k(n-1) + 1)$, we are able to identify the filtration of the last $s - \tau + 1$ cones

$$C_\tau \subseteq \ldots \subseteq C_s = \mathcal{P}_{n+1,2d} \tag{2}$$

with the $(n, 2d)$ case. This ensures the elimination of one variable in an inductive argument. Repeating that consideration, we at last arrive in the base case $(2, 2d)$. This is a Hilbert case and therefore fully understood. It thus remains to investigate the situation of the first $\tau + 1$ cones

$$\Sigma_{n+1,2d} = C_0 \subseteq \ldots \subseteq C_\tau. \tag{3}$$

Indeed, putting (2) and (3) together recovers the initial filtration (1).

In (3), an immediate application of the main result from Blekherman et al. (2016) allows us to conclude that the SOS cone always coincides with C_1. Furthermore, a slight variation of this result ensures $\Sigma_{n+1,2d} = C_i$ for any $i \in \{1, \ldots, n\}$. Thus,

$$\Sigma_{n+1,2d} = C_0 = \ldots = C_n. \tag{4}$$

After that, for inclusions of the type $C_i \subseteq C_{i+1}$ with $i \in \{n, \ldots, \tau - 1\}$, the situation is more evolved and other methods have to be applied.

For example, in the basic non Hilbert case $(4, 4)$, exactly two strictly separating intermediate cone in (3) are identifiable. Indeed, the famous *Robinson form*

$$R(X_0, X_1, X_2, X_3) := X_0^2(X_0 - X_3)^2 + X_1^2(X_1 - X_3)^2 + X_2^2(X_2 - X_3)^2 + 2X_0X_1X_2(X_0 + X_1 + X_2 - 2X_3)$$

and the *Choi-Lam form*

$$W(X_0, X_1, X_2, X_3) := X_0^2X_1^2 + X_0^2X_2^2 + X_1^2X_2^2 + X_3^4 - 4X_0X_1X_2X_3$$

both certify the proper containment $\Sigma_{4,4} \subsetneq \mathcal{P}_{4,4}$ (see Robinson (1969) and Choi and Lam (1977a,b)). In particular, the Robinson form was alongside the *Motzkin form* (see Motzkin (1967)) one of the first forms found separating the SOS and PSD cone in a basic non Hilbert case. Both were firstly mentioned roughly nine decades after Hilbert's original abstract proof from 1888 in the late 1960's. The Choi-Lam form followed in 1977.
Now, a deeper reaching investigation of the Robinson form, the Choi-Lam form and a variation of the Choi-Lam form under permuation of variables reveals

$$C_3 \subsetneq C_4 \subsetneq C_5 \subsetneq C_6 \tag{5}$$

in the basic non Hilbert case of quaternary quartics. Furthermore,

$$\Sigma_{4,4} = C_0 = C_1 = C_2 = C_3$$

by (4) and $C_6 \subseteq \ldots \subseteq C_{10} = \mathcal{P}_{4,4}$ corresponds to (2) and with that to the $(3, 4)$ case by our inductive argument. The ternary quartics describe a Hilbert case and, consequently, the latter subfiltration collapses to

$$C_6 = \ldots = C_{10} = \mathcal{P}_{4,4}.$$

Putting it all together, we thus fully understand the situation in the basic non Hilbert case of quaternary quartics.
In particular, we strengthen Hilbert's original observation from 1888 in the quaternary quartics case by testifying the existence of two distinct strictly separating intermediate cones between the SOS and the PSD cone.

REFERENCES

Blekherman, G., Smith, G.G., and Velasco, M. (2016). Sums of squares and varieties of minimal degree. *J. Amer. Math. Soc.*, 29, no. 3, 893–913.

Choi, M. and Lam, T. (1977a). Extremal positive semidefinite forms. *Math. Ann.*, 231, 1–18.

Choi, M. and Lam, T. (1977b). An old question of Hilbert. *Queen's Pap. Pure Appl. Math.*, 46, 385–405.

Choi, M., Lam, T., and Reznick, B. (1995). Sums of squares of real polynomials. In *K-Theory and Algebraic Geometry: Connections with Quadratic Forms and Division Algebras, Part 2*, volume 58.2 of *Proceedings of Symposia in Pure Mathematics*, 103–126. American Mathematical Society. doi: 10.1090/pspum/058.2/1327293.

Goel, C. (2020). Cones in and around the sums of squares cone. In *Real algebraic geometry with a view toward Hyperbolic Programming and Free Probability*, Oberwolfach Rep. 17, 639–712. doi:10.4171/OWR/2020/12.

Hilbert, D. (1888). *Über die Darstellung definiter Formen als Summe von Formenquadraten*, volume 32. doi: 10.1007/BF01443605.

Marshall, M. (2008). Positive polynomials and sum of squares. *Math. Surveys & Monographs*, 146.

Motzkin, T. (1967). The arithmetic-geometric inequalities. In *Inequalities (Proc. Sympos. Wright-Patterson Air Force Base, Ohio, 1965)*. Academic Press.

Plaumann, D. (2020). *Einführung in die Algebraische Geometrie (German)*, volume 1. Springer-Verlag.

Powers, V. and Wörmann, T. (1998). An algorithm for sums of squares of real polynomials (English summary). *J. Pure Appl. Algebra*, 127, no. 1, 99–104.

Robinson, R.M. (1969). Some definite polynomials which are not sums of squares of real polynomials. In *Selected questions of algebra and logic (a collection dedicated to the memory of A. I. Mal'cev)*, volume 16, 554. "Nauka" Sibirsk. Otdel. Novosibirsk.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Deterministic optimal control problem on Riemannian manifolds under probability knowledge of the initial condition

F. Jean * **O. Jerhaoui** ** **H. Zidani** ***

* *UMA, ENSTA Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France (e-mail: frederic.jean@ensta-paris.fr).*
** *UMA, ENSTA Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France (e-mail: othmane.jerhaoui@ensta-paris.fr).*
*** *INSA Rouen Normandie, 76800 Saint-Étienne-du-Rouvray, France (e-mail: hasnaa.zidani@insa-rouen.fr).*

Abstract: This paper concerns an optimal control problem on the space of probability measures over a compact Riemannian manifold. The motivation behind it is to model certain situations where the central planner of a deterministic controlled system has only a probabilistic knowledge of the initial condition. The lack of information here is very specific. In particular, we show that the value function verifies a dynamic programming principle and we prove that it is the unique viscosity solution to a suitable Hamilton Jacobi Bellman equation. The notion of viscosity is defined using test functions that are directionally differentiable in the the space of probability measures.

Keywords: Optimal Control, Viscosity solutions, Hamilton Jacobi Bellman equation, Wasserstein spaces, Multi-agent systems.

1. INTRODUCTION

The study of optimal control problems and viscosity theory in the space of probability measures has been an active area of research in the mathematical community in the last years, in particular because of its potential real-world applications in the modeling of multi-agent systems. The potential applications include crowd dynamics modeling, opinion formation process modeling, herd analysis, social network analysis, autonomous multi-vehicle navigation and the modeling of uncertainties on the initial state of a deterministic controlled system.

At the individual level, the behavior of each agent is dictated not only by local interactions but also by the *non local* interactions that depend on the distribution of all agents. When the number of agents is assumed to be very large, the complexity of the system grows extremely fast. A suitable way to modelize this problem is through a macroscopic approach, where the discrete collection of agents is replaced by a spatial density that evolves in time. If we assume further that the total number of agents remains constant at all times during the evolution of the system, then one can normalize the density of the agents and assume that its total mass is equal to 1.

Hence, the evolution of the multi-agent system, seen as normalized spatial density in a given base space X (typically the Euclidean space or a Riemannian manifold), is described by a curve $t \mapsto \mu_t \in \mathcal{P}(X)$, where $\mathcal{P}(X)$ is the space of Borel probability measures over X, and μ_t represents the spatial density of the multi-agent system at a given time $t \geq 0$. The conservation of the mass along the trajectory $t \mapsto \mu_t$ is described by the following continuity equation

$$\partial_t \mu_t + div(v_t \mu_t) = 0,$$

where $v_t(.)$ is a time-dependent Borel vector field, and the equation is understood in the sense of distributions.

In this paper, we take the base space $X = M$ to be a compact Riemannian manifold without boundary. We propose to study a simple model of multi-agent systems, where the non local interactions between the agents are not considered. This problem can be interpreted as a deterministic control system with imperfect information on the initial condition, i.e. the initial condition is not known precisely by the controller, but they only know that the initial condition follows a probability distribution $\mu_0 \in \mathcal{P}(M)$. More precisely, consider the following controlled equation:

$$\begin{cases} \dot{Y}(t) = f(Y(t), u(t)), \quad t \in [t_0, T], \\ Y(t_0) = x_0, \quad u(t) \in U, \end{cases} \tag{1}$$

where $f : M \times U \to TM$ is the dynamics, assumed to be Lipschitz with respect to the first variable and continuous with respect to the second variable, $x_0 \in M$ and $t_0 \in [0, T]$. The set U is the set of admissible control values which is assumed to be a compact subset of some metric space. the control function $u(.) \in U$ is a Borel measurable function $u : [t_0, T] \to U$. The main feature of this problem is that the initial position x_0 is not perfectly known, but rather distributed along the probability measure μ_0. Notice that, since $f(., u(t))$ is Lipschitz continuous and bounded, the evolution curve of the uncertainty, $t \mapsto \mu_t$ starting from μ_0, is the unique solution to the equation

$$\begin{cases} \partial_t \mu_t + div(f(., u(t))\mu_t) = 0, \ t \in (t_0, T), \\ \mu_{t_0} = \mu_0, \end{cases}$$

in the distributional sense. The measures μ_t are obtained by the pushforward of μ_0 by the flow at time t of the controlled equation (1).
The controller aims at minimizing the following final cost:

$$L(\mu) = \int l(y)d\mu(y),$$

where $l : M \to \mathbb{R}$ is a Lipschitz function. An immediate consequence of this assumption is that the function $L : \mathcal{P}(M) \to \mathbb{R}$ inherits the Lipschitz property from l as well. The quantity $L(\mu_T)$ represents the expectation of the deterministic final cost with respect to the measure μ_T.

To this optimal control problem, we associate the following value function:

$$\vartheta(t_0, \mu_0) = \inf_{u(.)\in U} L(\mu_T).$$

The first main goal of this paper is to study the properties and the regularity of the value function. In particular we will show that the value function is Lipschitz continuous with respect to both variables and that it verifies the dynamic programming principle. The second goal of the paper is to prove that value function can be characterized as the unique viscosity solution of a suitable Hamilton Jacobi Bellman equation (HJB in short) of the form

$$\begin{cases} \partial_t v + H(\mu, D_\mu v) = 0, \quad (t,\mu) \in [0,T) \times \mathcal{P}_2(M), \\ v(T,\mu) = L(\mu), \end{cases}$$

in the space of probability measures $\mathcal{P}(M)$. We aim at transposing the viscosity theory techniques that are used in the classical theory (Crandall et al. (1992)) to the space of Borel probability measures $\mathcal{P}_2(M)$. In particular, we define a suitable notion of viscosity using a class of real valued functions that admit *directional derivatives* at all points $\mu \in \mathcal{P}(M)$. We then prove a local comparison principle between any bounded upper semicontinuous subsolution and any lower semicontinuous supersolution. Finally, we prove that the value function is the unique viscosity solution to the above HJB equation by using the dynamic programming principle verified by the value function.
This paper is expository. All the results asserted in here are proven in Jean et al. (2022). The paper is structured as follows. In Section 2, we recall some classical notions of optimal transport theory and the geometry of the space of probability measures. In Section 3, we formulate the Mayer problem in the space of probability measures and we give the main properties of the value functions. Section 4 is devoted to the study of a suitable HJB equation that characterizes the value function. In particular, we define the Hamiltonian we are going to work with, then we define a notion of viscosity using a class of test functions that are directionally differentiable, we show the comparison principle and we show that the value function is the unique viscosity of the HJB equation.

2. PRELIMINARIES

In this section, we recall some facts about optimal transport and the geometry of Wasserstein spaces. Let $(M, \langle ., . \rangle)$ be a finite dimensional, compact and connected Riemannian manifold without boundary. We denote by $|\,.\,|$ the associated norm on the tangent bundle TM, and by $d(.,.)$ its Riemannian distance on M. The metric space (M, d), is a complete, separable and compact space and its topology is equivalent to the topology of the differentiable manifold M. The tangent bundle TM is itself a complete and separable Riemannian manifold when endowed with the Sasaki metric (Sasaki (1962)). We denote by $d_{TM}(.,.)$ the Riemannian distance on TM associated with the Sasaki metric.

2.1 The Wasserstein space $\mathcal{P}_2(M)$

We denote by $\mathcal{P}(M)$ the set of Borel probability measures over M and $\mathcal{P}_2(M)$ the set of Borel probability measures with bounded second moments:

$$\mathcal{P}_2(M) := \{\mu \in \mathcal{P}(M) : \int d^2(x, x_0)d\mu(x) < \infty, \ \forall x_0 \in M\}$$

Actually, since M is compact, we have $\mathcal{P}_2(M) = \mathcal{P}(M)$ but we will keep using the notation $\mathcal{P}_2(M)$. Recall that for any two topological spaces X and Y, any Borel probability measure μ on X and any Borel function $g : X \to Y$, the pushforward measure $g\sharp\mu$ on Y is defined by

$$g\sharp\mu(A) = \mu(g^{-1}(A)) \quad \forall A \subset Y, \text{ a Borel set},$$

or equivalently, for all $h : Y \to \mathbb{R}$, a Borel measurable and bounded function, we have:

$$\int h \, dg\sharp\mu = \int h \circ g \, d\mu.$$

We define the Wasserstein distance $W_2(.,.)$ over $\mathcal{P}_2(M)$ by

$$W_2(\mu,\nu) := \sqrt{\inf\Big\{ \int d^2(x,y)d\gamma(x,y) \Big\}},$$

where the infimum is taken over all Borel probability measures of $M \times M$ that have marginals μ and ν, i.e. for all $A, B,$ Borel sets of M, we have

$$\gamma(A \times M) = \mu(A) \quad \text{and} \quad \gamma(M \times B) = \nu(B).$$

The metric space $(\mathcal{P}_2(M), W_2)$ is complete and separable. Furthermore, it is a geodesic space, i.e. any two points of $\mathcal{P}_2(M)$ can be joined by at least one geodesic. We recall that a curve $\alpha : [0,1] \to \mathcal{P}_2$ is a geodesic if

$$W_2(\alpha_t, \alpha_s) \leq |t-s| W_2(\alpha_0, \alpha_1), \ \forall t, s \in [0,1].$$

2.2 The tangent space $T_\mu(\mathcal{P}_2(M))$

In this subsection, we will adopt a purely metric perspective to define the tangent space of $(\mathcal{P}_2(M), W_2)$ at a given point μ. First, Let $\mathcal{P}(TM)$ be the set of Borel probability measures over TM. Since (TM, d_{TM}) is a complete geodesic space, we can define the Wasserstein space over TM as the set of all $\eta \in \mathcal{P}(TM)$ such that

$$\int d^2_{TM}\Big((x,v),(x_0,v_0)\Big)d\eta(x,v) < \infty,$$

for all $(x_0, v_0) \in TM$. We denote it by $\mathcal{P}_2(M)$. We endow it with the usual Wasserstein distance for any $\eta, \gamma \in \mathcal{P}_2(TM)$:

$$W_2(\gamma,\eta) := \sqrt{\inf\Big\{ \int d^2_{TM}(x,y)d\beta(x,y) \Big\}},$$

the infinimum is taken over all admissible plans β with marginals γ and η. We are now able to define the tangent space at a point $\mu \in \mathcal{P}_2(M)$.

Definition 1. (Tangent space). Let $\mu \in \mathcal{P}_2(M)$. The tangent space $T_\mu(\mathcal{P}_2(M)) \subset \mathcal{P}_2(TM)$, is the set of plans $\gamma \in \mathcal{P}_2(TM)$ such that $\pi^M\sharp\gamma = \mu$, where $\pi^M : TM \to M$ is the canonical projection onto M.

The tangent space at a point μ has a geometric meaning. In fact, it encodes all the information about geodesics emanating from μ as we describe hereafter.
Let $\exp : TM \to M$ be the exponential map of $(M, \langle ., . \rangle)$. The exponential $\mathbf{exp}_\mu(\gamma)$ of a plan $\gamma \in T_\mu(\mathcal{P}_2(M))$ is defined by

$$\mathbf{exp}_\mu(\gamma) := \exp \sharp \gamma \in \mathcal{P}_2(M).$$

We define the map $\mathbf{exp}_\mu^{-1} : \mathcal{P}_2(M) \to T_\mu(\mathcal{P}_2(M))$ by

$$\mathbf{exp}_\mu^{-1}(\nu) := \{ \gamma \in T_\mu(\mathcal{P}_2(M)) \ : \ \mathbf{exp}_\mu(\gamma) = \nu \text{ and } \int |v|^2 \, \mathrm{d}\gamma(x,v) = (d_W(\mu,\nu))^2 \},$$

or in other words, the set of plans $\gamma \in \mathcal{P}_2(TM)$ such that $(\pi^M, \exp)\sharp\gamma$ is an optimal plan from μ to ν and $\int |v|^2 \, \mathrm{d}\gamma(x,v) = (d_W(\mu,\nu))^2$. We introduce the following notation

$$\Delta_t(x,v) = (x, tv), \quad \forall t \in \mathbb{R}, \ (x,v) \in TM.$$

Remark 2. The map $\mathbf{exp}_\mu^{-1}$ is not really an inverse map to $\mathbf{exp}_\mu$ since only optimal plans in the inverse image of ν are considered. While this might seem confusing, the map $\mathbf{exp}_\mu^{-1}$ is defined this way so that for all $\gamma \in \mathbf{exp}_\mu^{-1}(\nu)$, the curve $t \to \exp(\Delta_t)\sharp\gamma$ is a geodesic connecting μ and ν, see the theorem below.

Theorem 3. (Gigli, 2011, Theorem 1.11) Let $\mu, \nu \in \mathcal{P}_2(M)$. A curve $(\mu_t) : [0,1] \to \mathcal{P}_2(M)$ is a geodesic connecting μ to ν if and only if there exists a plan $\gamma \in \mathbf{exp}_\mu^{-1}(\nu)$ such that

$$\mu_t := \exp \circ \Delta_t \sharp \gamma, \quad \forall \, t \in [0,1]. \tag{2}$$

The plan γ is uniquely identified by the geodesic. Moreover, for any $t \in (0,1)$ there exists a unique optimal plan from μ to μ_t. Finally, if there exist two different geodesics connecting μ to ν, they do not intersect in intermediate times (i.e. on $(0,1)$).

Introducing the following rescaling of a plan:

$$t \cdot \gamma = \Delta_t \sharp \gamma, \quad \forall t \in \mathbb{R}, \ \gamma \in \mathcal{P}_2(TM),$$

equation (2) can be rewritten in a more elegant way as

$$\mu_t = \exp \circ \Delta_t \sharp \gamma = \mathbf{exp}_\mu(t \cdot \gamma), \quad \forall t \in [0,1].$$

With the characterization of geodesics in Theorem 3, notice that for any $\mu \in \mathcal{P}_2(M)$, all plans $\gamma \in T_\mu(\mathcal{P}_2(M))$ that produce geodesics, i.e. such that

$$t \mapsto \mathbf{exp}_\mu(t \cdot \gamma), \tag{3}$$

is a geodesic defined in some right neighborhood of 0, say $[0, \varepsilon]$, can be seen as *initial velocities* of these geodesics. We mention that not all curves of this form are necessarily geodesics but all geodesics are of this form.
Moreover, using this characterization of geodesics, we can define a class of real valued functions $f : \mathcal{P}_2(M) \to \mathbb{R}$ that are *directionally differentiable* along all geodesics. In particular, the squared Wasserstein distance function

$$\mu \mapsto W^2(\mu, \sigma),$$

with $\sigma \in \mathcal{P}_2(M)$ fixed, is directionally differentiable along all geodesics. In fact, a much more general result holds: the squared Wasserstein distance is directionally differentiable along all curves of the form (3) even though they are not geodesics. For more details on this, we refer to Gigli (2011). We will only give the following result for the squared Wasserstein distance, which we will need in order to define test functions for viscosity notion in Section 4.

Theorem 4. (Gigli, 2011, Theorem 4.2) Let $\mu, \sigma \in \mathcal{P}_2(M)$, and $g : M \to TM$ be a squared integrable vector field with respect to μ. Let $\gamma = g\sharp\mu \in T_\mu(\mathcal{P}_2(M))$ and $t \mapsto \mathbf{exp}_\mu(t \cdot \gamma)$ be a curve starting from μ, *not necessarily a geodesic*. Then it holds

$$\frac{d}{dt}\Big|_{t=0} W_2^2(\mathbf{exp}_\mu(t \cdot \gamma), \sigma)^2 = -2 \sup \int \langle g(x), v \rangle d\zeta(x,v),$$

where the supremum is taken over all $\zeta \in \mathbf{exp}_\mu^{-1}(\sigma)$. We denote it by

$$D_\mu W_2^2(., \sigma) \cdot (g\sharp\mu) := \frac{d}{dt}\Big|_{t=0} W_2^2(\mathbf{exp}_\mu(t \cdot \gamma), \sigma)^2,$$

and is understood as the *differential* of $W_2^2(., \sigma)$ along $g\sharp\mu$.

3. OPTIMAL CONTROL PROBLEM IN $\mathcal{P}_2(M)$

Let $T > 0$ and U be a compact subset of a metric space. Consider the dynamics, defined for $T > t_0 \geq 0$ and $x_0 \in M$, as

$$\begin{cases} \dot{Y}(t) = f(Y(t), u(t)), & t \in [t_0, T], \\ Y(t_0) = x_0, u(t) \in U, \end{cases} \tag{4}$$

where $f : M \times U \to TM$ satisfies the following assumptions:

$$\textbf{(H)} \begin{cases} f : M \times U \to TM \text{ is continuous and Lipschitz} \\ \text{ continuous with respect to the state, i.e.} \\ \exists k > 0 : d_{TM}(f(x,u), f(y,u)) \leq k \, d(x,y), \\ \forall u \in U, \ (x,y) \in M \times M. \end{cases}$$

$\textbf{(H)}_{co}$: for all $x \in M$, the set
$f(x,U) := \{f(x,u) \ : \ u \in U\}$ is convex.

We define the set of open-loop controls by

$$\mathcal{U} := \{u : [0,T] \to U \ : \ u(.) \text{ is measurable}\}.$$

The control problem aims at minimizing the final cost

$$\int l(Y_T^{t_0,x_0,u}) \, d\mu_0(x_0),$$

over all trajectories that are solutions of the dynamics (4) with the initial condition $x_0 \in M$, distributed along the measure $\mu_0 \in \mathcal{P}_2(M)$. We consider the following assumption:

$\textbf{(H}_\textbf{l}\textbf{)}$ $l : M \to \mathbb{R}$ is Lipschitz continuous with constant $Lip(l)$.

When μ_0 is equal to the Dirac mass δ_{x_0}, the resulting system corresponds to the classical case without uncertainties on the initial condition. This problem is thoroughly studied in the literature. When μ_0 is any probability measure of $\mathcal{P}_2(M)$, it is better to see this problem as an optimal control problem defined on the space of Borel probability measures $\mathcal{P}_2(M)$. First we rewrite the final cost the following way

$$\int l(Y_T^{t_0,x_0,u}) \, d\mu_0(x_0) = \int l(y) \, dY_T^{t_0,.,u}\sharp\mu_0 \, (y),$$

and we minimize this cost over the set of trajectories $t \mapsto \mu_T^{t_0,\mu_0,u}$ of the space $\mathcal{P}_2(M)$ that verify

$$\begin{cases} \mu_t^{t_0,\mu_0,u} = Y_t^{t_0,.,u}\sharp\mu_0, \quad t \in [t_0,T], \text{ and } x \mapsto Y_t^{t_0,x,u} \\ \text{is the flow of (4)}, \\ \mu_{t_0} = \mu_0. \end{cases}$$

Since $x \mapsto Y_t^{t_0,x,u} \in M$ and $\mu_0 \in \mathcal{P}_2(M)$, then $t \mapsto Y_t^{t_0,.,u}\sharp\mu_0 \in \mathcal{P}_2(M)$ for all $t \in [t_0,T]$. It is a known

fact, Ambrosio et al. (2008); Bernard (2008), that each trajectory $t \mapsto \mu_t^{\mu_0,u}$ is the unique solution to the following continuity equation

$$\begin{cases} \partial_t \mu_t^{t_0,\mu_0,u} + div(f(.,u(t))\mu_t^{t_0,\mu_0,u}) = 0,\ t \in (t_0,T) \\ \mu_{t_0}^{\mu_0,u} = \mu_0. \end{cases} \tag{5}$$

In the distributional sense. Hence the optimal control problem can be rewritten in the following way:

$$\begin{cases} \min\limits_{u(.)\in U} \int l(y) d\mu_T^{t_0,\mu_0,u}(y) = L(\mu_T^{t_0,\mu_0,u}), \\ \text{such that } \begin{cases} \partial_t \mu_t^{t_0,\mu_0,u} + div(f(.,u(t))\mu_t^{t_0,\mu_0,u}) = 0, \\ \mu_{t_0}^{\mu_0,u} = \mu_0,\ t \in (t_0,T). \end{cases} \end{cases} \tag{6}$$

The associated *value function* to the above optimal control problem is defined as

$$\vartheta(t_0,\mu_0) := \inf_{u(.)\in\mathcal{U}} \int l(y)\, d\mu_T^{t_0,\mu_0,u}(y) = L(\mu_T^{t_0,\mu_0,u}).$$

Under hypotheses $(\boldsymbol{H})$, $(\boldsymbol{H_l})$ and $(\mathbf{H})_{co}$, we have the following two properties of the value function.

Proposition 5. (Jean et al. (2022)). Assume $(\boldsymbol{H})$, $(\boldsymbol{H_l})$ and $(\mathbf{H})_{co}$. Then, the value function ϑ is Lipschitz continuous on $[0,T] \times \mathcal{P}_2(M)$. In particular, ϑ is bounded.

Theorem 6. (Jean et al. (2022)). Let $\mu \in \mathcal{P}_2(M)$, $t \in [0,T]$ and $h \in [t, T-t]$. Then it holds

$$\vartheta(t,\mu) = \inf_{u\in\mathcal{U}} \vartheta(t+h, \mu_{t+h}^{t,\mu,u}).$$

4. HJB EQUATION IN $\mathcal{P}_2(M)$

The Hamiltonian we will work with has the following expression:

$$H(\mu, D_\mu v) = \inf_{u\in U} D_\mu v \cdot \Big(f(.,u)\sharp\mu\Big), \tag{7}$$

with $v : \mathbb{R} \times \mathcal{P}_2(M) \to \mathbb{R}$ is a real valued function that admits directional derivatives along the time variable and the measure variable. The function v represents a test function that has the following form:

$$\forall (t,\mu) \in [0,T) \times \mathcal{P}_2(M),\ v(t,\mu) = \psi(t) + a\, W_2^2(\mu,\sigma),$$

with $a \in \mathbb{R}$ and $\sigma \in \mathcal{P}_2(M)$ fixed and $\psi : [0,T) \to \mathbb{R}$ is a continuously differentiable function, $\mu \mapsto W_2^2(\mu,\sigma)$ is directionally differentiable in the sense of Theorem 4. We consider the following Hamilton Jacobi Bellman equation:

$$\begin{cases} \partial_t v + H(\mu, D_\mu v) = 0, \quad (t,\mu) \in [0,T) \times \mathcal{P}_2(M), \\ v(T,\mu) = L(\mu). \end{cases} \tag{8}$$

Next, we define the test functions that we are going to use to define the notion of viscosity solutions.

Definition 7. (Test functions).
Let $\mathcal{TEST}_1$ be the set defined as:

$$\mathcal{TEST}_1 := \{(t,\mu) \to \psi(t) + a(\,(d_W(\mu,\sigma))^2\,) : a \in \mathbb{R}^+, \sigma \in \mathcal{P}_2(M)\},$$

where $\psi : [0,T) \to \mathbb{R}$ is a C^1 function.
We set $\mathcal{TEST}_2 = -\mathcal{TEST}_1 := \{-\phi : \phi \in \mathcal{TEST}_1\}$.

Definition 8. (Viscosity solutions).

- We say that a function $v : [0,T) \times \mathcal{P}_2(M) \to \mathbb{R}$ satisfies the inequality
$$\partial_t v + H(\mu, D_\mu v) \geq 0,$$
at $(t,\mu) \in [0,T) \times \mathcal{P}_2(M)$ in the viscosity sense if v is upper semicontinuous and for all $\mathcal{TEST}_1$ functions $\phi : [0,T) \times \mathcal{P}_2(M) \to \mathbb{R}$ such that $v - \phi$ attains a maximum at (t,μ), we have
$$\partial_t \phi + H(\mu, D_\mu \phi) \geq 0.$$
A function v satisfying $\partial_t v + H(\mu, D_\mu v) \geq 0$ on $[0,T) \times \mathcal{P}_2(M)$ in the viscosity sense is called a *viscosity subsolution* of (8).
- Similarly, we say that a function $v : [0,T) \times \mathcal{P}_2(M) \to \mathbb{R}$ satisfies the inequality
$$\partial_t v + H(\mu, D_\mu v) \leq 0,$$
at $(t,\mu) \in [0,T) \times \mathcal{P}_2(M)$ in the viscosity sense if v is lower semicontinuous and for all $\mathcal{TEST}_2$ functions $\phi : [0,T) \times \mathcal{P}_2(M) \to \mathbb{R}$ such that $v - \phi$ attains a minimum at (t,μ), then
$$\partial_t \phi + H(\mu, D_\mu \phi) \leq 0.$$
A function v satisfying $\partial_t v + H(\mu, D_\mu v) \leq 0$ on $[0,T) \times \mathcal{P}_2(M)$ in the viscosity sense is called a *viscosity supersolution* of (8).
- We say that a continuous function $v : [0,T] \times \mathcal{P}_2(M) \to \mathbb{R}$ is a *viscosity solution* of (8) if it is both a supersolution and a subsolution on $[0,T) \times \mathcal{P}_2(M)$ and verifies
$$v(T,\mu) = L(\mu).$$

Theorem 9. (Jean et al. (2022)). Assume $(\boldsymbol{H})$ and $(\boldsymbol{H_l})$. Let $u, v : [0,T] \times \mathcal{P}_2(M) \to \mathbb{R}$ be respectively bounded upper semicontinuous subsolution and lower semicontinuous supersolution on $[0,T] \times \mathcal{P}_2(M)$. Then it holds:

$$\sup_{[0,T]\times\mathcal{P}_2(M)} (v-w)_+ \leq \sup_{\{T\}\times\mathcal{P}_2(M)} (v-w)_+,$$

where $(a)_+ = max(a,0)$.

Theorem 10. (Jean et al. (2022)). Assume $(\boldsymbol{H})$, $(\boldsymbol{H_l})$ and $(\boldsymbol{H_{co}})$. Then the value function ϑ is the unique continuous viscosity solution to (8).

REFERENCES

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Bernard, P. (2008). Young measures, superposition and transport. *Indiana University mathematics journal*, 247–275.

Crandall, M.G., Ishii, H., and Lions, P.L. (1992). User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American mathematical society*, 27(1), 1–67.

Gigli, N. (2011). On the inverse implication of brenier-mccann theorems and the structure of (p 2 (m), w 2). *Methods and Applications of Analysis*, 18(2), 127–158.

Jean, F., Jerhaoui, O., and Zidani, H. (2022). Deterministic optimal control on riemannian manifolds under probability knowledge of the initial condition. *Unpublished*.

Sasaki, S. (1962). On the differential geometry of tangent bundles of riemannian manifolds ii. *Tohoku Mathematical Journal, Second Series*, 14(2), 146–155.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A Priori Parameter Identifiability in Complex Reaction Networks

Ragini Sreenath [*] **Sridharakumar Narasimhan** [*,***,****]
Nirav Bhatt [**,***,****]

[*] *Department of Chemical Engineering*
[**] *Department of Biotechnology*
[***] *Robert Bosch Center for Data Science and Artificial Intelligence*
[****] *pCoE for Network Systems Learning, Control and Evolution*
Indian Institute of Technology, Madras, Chennai–600036, India,
(e-mail: ragini.pvs06@gmail.com, 147199 sridharkrn@iitm.ac.in, and niravbhatt@iitm.ac.in)

Abstract: Differential algebra-based theory and software have been widely used to study the *a priori* structural identifiability of nonlinear systems. This technique however fails to provide definitive answers for complex reaction networks which involve several reactions and species. In this work, for reaction systems following mass action kinetics, using the theory of reaction extents, we show that identifiability can be ascertained by determining the rank of a matrix. Further, we show that for systems involving most bi-molecular reactions, the parameters are guaranteed to be identifiable, if R (where R = number of independent reactions) species that satisfy a rank condition are measured.

Keywords: Parameter Identifiability, Ritts Pseudo–Division Algorithm, Extent Domain, Complex Reaction Networks, Systems Biology

1. INTRODUCTION

It is important to identify reliable and accurate models of reaction systems for use in model-based analysis tasks such as simulation, control, and optimisation. In practice, the identification of models is an iterative process involving: (i) experimental data generation using carefully planned experiments, and (ii) fitting a proposed model (or a set of proposed models) to generated data (van Riel, 2006). It is important to investigate whether the unknown parameters can be uniquely estimated from observed data before investing resources, time and effort in performing actual experiments (Chis et al., 2011).

A priori structural identifiability addresses the issue of whether it is possible to recover the unknown parameters uniquely from a proposed model structure from error-free data (Ljung and Glad, 1994). The problem of structural rate identifiability in reaction systems involves determining the actual parameters appearing in the kinetic rate laws. This has been studied in the larger context of parameter identifiability of ODE models and different methods based on the Taylor series, generating functions, differential algebra, implicit function theorem, etc. have been proposed in the existing literature and have been compared by Chis et al. (2011). Differential algebra-based methods involve two steps. In the first step, a set of differential polynomials called the characteristics set that relates the inputs, outputs, and their derivatives is derived. This is followed by testing for the injectivity of the coefficient maps of the differential polynomials. A user-friendly software tool, DAISY (Differential Algebra for Identifiability of SYstems) (Bellu et al., 2007; Saccomani et al., 2003) to check identifiability in a specific class of nonlinear systems that contain polynomial or rational functions has been developed. Systems described by non-rational functions can be approximated by rational functions and we have shown that the differential-algebraic methods can provide a partial answer to the question of structural identifiability (Jain et al., 2019). Although the algorithms are guaranteed to converge, the worst-case time complexity is very large. Hence, methods based on DAISY may fail to provide a definitive result in reasonable time (Saccomani et al., 2010; Varghese et al., 2018). Moreover, the methods do not exploit the underlying structure of the system, e.g., reaction networks which is the focus of this work.

In the previous work, we applied the theory of reaction variants and in-variants (Amrhein et al., 2010) to simplify the problem of determining structural rate identifiability in reaction systems [1] (Varghese et al., 2018). For a class of reaction systems, an appropriate extent of reaction was used to develop an alternate, but equivalent state-space model with significantly lower dimensions. The advantages of the method were demonstrated in situations where the use of conventional methods and software e.g., DAISY failed. This was also extended to the situation when the number of participating reactions is more than the number of independent reactions, and hence applicable to complex reaction networks which involve a large number of reactions and species.

In this work, for reaction systems following mass-action kinetics, we show that the characteristic set is readily

[1] In the interest of brevity, we refer to structural rate identifiability as parameter identifiability or simply identifiability

obtained by solving a set of linear equations alone. Identifiability can be ascertained by determining the rank of a matrix. Further, we show that for systems involving most bimolecular reactions, the parameters are guaranteed to be identifiable, if R (R = number of independent reactions) species that satisfy a rank condition are measured. This is significant because in many cases, this is possible by inspection alone. Hence, the need to perform computer algebra (or symbolic computation) required for differential-algebra methods and software such as DAISY can be avoided.

2. PRELIMINARIES

2.1 Differential Algebra approach to identifiability

Consider a nonlinear system:

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{p}), \qquad \mathbf{x}(0) = \mathbf{x}_0, \\ \mathbf{y} &= \mathbf{g}(\mathbf{x}, \mathbf{u}, \mathbf{p}) \end{aligned} \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{R}^S$ is the state vector, $\mathbf{u} \in \mathbb{R}^\upsilon$ is the input vector, $\mathbf{y} \in \mathbb{R}^m$ is the output vector, and $\mathbf{p} \in \mathbb{R}^p$ is the parameter vector. It is assumed that $\mathbf{f}(\cdot) : \mathbb{R}^S \to \mathbb{R}^S$ and $\mathbf{g}(\cdot) : \mathbb{R}^S \to \mathbb{R}^m$ are Lipschitz continuous with respect to $\mathbf{x}$, and $\mathbf{u}$. When the nonlinear functions in Eq. (1) are rational or polynomial, identifiability of the parameters can be determined as follows. A set of m differential polynomials in terms of input variables $\mathbf{u}$, output variables $\mathbf{y}$, derivatives of inputs, outputs and parameters $\mathbf{p}$ are determined and denoted as input-output relations or characteristic set of the system (1). This set of input-output relations is determined from (1) using the Ritt's pseudo–division algorithm (Ritt, 1950; Saccomani et al., 2003). Let $\mathbf{h}(\mathbf{p})$ denote the vector of coefficients in the input output map. Then, we can check the injectivity of the coefficients map $\mathbf{h}(\mathbf{p})$ of the m differential polynomials for determining a priori parameter identifiability by evaluating $\mathbf{h}(\mathbf{p})$ for an arbitrary $\mathbf{p}^*$. The Büchberger algorithm is usually employed to obtain the Gröbner basis for this set of equations $\mathbf{h}(\mathbf{p}) = \mathbf{h}(\mathbf{p}^*)$. Depending on the nature of solutions of $\mathbf{h}(\mathbf{p}) = \mathbf{h}(\mathbf{p}^*)$, we can classify the system as follows.

Definition 1. (Globally Identifiable). The model (1) is globally identifiable from the input-output data if and only if for any arbitrary $\mathbf{p}^*$, $\mathbf{h}(\mathbf{p}) = \mathbf{h}(\mathbf{p}^*)$ has a unique solution $\mathbf{p} = \mathbf{p}^*$.

Definition 2. (Locally Identifiable). If there exists multiple but finite number of distinct solutions for $\mathbf{h}(\mathbf{p}) = \mathbf{h}(\mathbf{p}^*)$ then the model (1) is locally identifiable.

Definition 3. (Unidentifiable). If there are infinite number of solutions for $\mathbf{h}(\mathbf{p}) = \mathbf{h}(\mathbf{p}^*)$ then the model (1) is unidentifiable.

2.2 Models of Reaction Systems

We consider an isothermal constant volume (V_0) batch reaction system with S species and R reactions. $\mathbf{N}$ is the $R \times S$-dimensional stoichiometric matrix and $\mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta})$ is the R-dimensional vector of reaction rates, and $\boldsymbol{\theta}$ is the p-dimensional vector of parameters. A reaction rate is typically a nonlinear function of the concentrations $\mathbf{c}$ and the parameter vector $\boldsymbol{\theta}$. The independent reactions can be defined as follows (Bhatt, 2011):

Definition 4. (Independent reactions). R reactions are said to be independent if (i) the rows of $\mathbf{N}$ (stoichiometries) are linearly independent, i.e., rank $(\mathbf{N}) = R$, and (ii) there exists some finite time interval for which the reaction rate profiles $\mathbf{r}(t)$ are linearly independent, i.e., $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{r}(t) = 0 \Leftrightarrow \boldsymbol{\beta} = \mathbf{0}_R$.

Without loss of generality, we assume that the reactions are independent. The mole balance equations for this system can be written as:

$$\begin{aligned} \dot{\mathbf{n}}(t) &= V_0\mathbf{N}^{\mathrm{T}}\mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}), \quad \mathbf{n}(0) = \mathbf{n}_0 \\ \mathbf{c}(t) &= \frac{\mathbf{n}(t)}{V_0} \end{aligned} \tag{2}$$

where $\mathbf{n}$ and $\mathbf{c}$ are the S-dimensional vectors of the number of moles, and concentrations, respectively. Without loss of generality, it is assumed that the initial concentrations ($\mathbf{c}_0$) are known. The model (2) can be written in terms of the concentrations as follows:

$$\dot{\mathbf{c}}(t) = \mathbf{N}^{\mathrm{T}}\mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}), \quad \mathbf{c}(0) = \mathbf{c}_0 \tag{3}$$

In practice, only a subset of concentrations are measured. Let $\mathbf{c}$ be partitioned as: $\mathbf{c}^{\mathrm{T}} = [\mathbf{c}_m^{\mathrm{T}}\ \mathbf{c}_u^{\mathrm{T}}]$, where $\mathbf{c}_m$ and $\mathbf{c}_u$ denote the measured and unmeasured species concentration respectively. Then, the above model can be expressed in conventional state space form as follows

$$\begin{aligned} \dot{\mathbf{c}}(t) &= \mathbf{N}^{\mathrm{T}}\mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}) \\ \mathbf{c}_m &= [\mathbf{I}_m\ \mathbf{0}]\,\mathbf{c} \end{aligned} \tag{4}$$

Definition 5. (Reaction variant, and invariant). Any set of R linearly independent variables that evolves with time and depends on reaction rates constitutes a reaction variant set. Any set of $(S - R)$ linearly independent variables that do not change with time constitutes a reaction invariant set. A linear transformation of the concentrations in (4) can be defined as follows:

$$\begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_{iv} \end{bmatrix} = \begin{bmatrix} \mathbf{N}^{T\dagger} \\ \mathbf{Q}^{\mathrm{T}} \end{bmatrix} (\mathbf{c} - \mathbf{c}_0) \tag{5}$$

where $\mathbf{x}_r$ is the $R-$dimensional vector of the extents of reaction, and $\mathbf{x}_{iv}$ is $(S - R)-$dimensional vector of invariant states. The invariant states do not change with time. '$\dagger$' denotes the Moore-Penrose pseudo–inverse of the matrix. $\mathbf{Q}$ is an $S \times S - R$-dimensional matrix such that $\mathbf{NQ} = \mathbf{0}$. The concentrations can be related to the extents of reaction as follows:

$$\mathbf{c}(t) = \mathbf{N}^{\mathrm{T}}\mathbf{x}_r + \mathbf{c}_0 \tag{6}$$

Typically, not all species are measured. Given a subset of measurements, under certain conditions, the reaction extents and unmeasured concentrations can be reconstructed using the following proposition.

Proposition 6. Let the matrix $\mathbf{N}$, and the initial conditions $\mathbf{c}_0$ be known and, without loss of generality, let $\mathbf{N}$ and $\mathbf{c}$ be partitioned as: $\mathbf{N} = [\mathbf{N}_m\ \mathbf{N}_u]$ and $\mathbf{c}^{\mathrm{T}} = [\mathbf{c}_m^{\mathrm{T}}\ \mathbf{c}_u^{\mathrm{T}}]$. Furthermore, let $\mathbf{c}_m(t)$ be measured without errors. If (i) rank $(\mathbf{N}_m) = R$, then the unmeasured concentrations $\mathbf{c}_u(t)$ can be reconstructed from the available $\mathbf{c}_m(t)$ in two steps as follows: (i) computation of the extents of reaction, $\mathbf{x}_r(t) = (\mathbf{N}_m^{\mathrm{T}})^\dagger(\mathbf{c}_m(t) - \mathbf{c}_{0,m})$, and (ii) reconstruction of the unmeasured concentrations $\mathbf{c}_u(t)$: $\mathbf{c}_u(t) = \mathbf{N}_u^{\mathrm{T}}\mathbf{x}_r(t) + \mathbf{c}_{0,u}$ (See Proof in (Bhatt, 2011))

Thus, the system can be defined in terms of a lower dimensional state space, viz., the R-dimensional space of

reaction variants or extents as follows:

$$\begin{aligned} \dot{\mathbf{x}}_r &= \mathbf{r}(\mathbf{x}, \boldsymbol{\theta}), & \mathbf{x}_r(0) &= \mathbf{0}, \\ \dot{\mathbf{x}}_{iv} &= \mathbf{0}, & \mathbf{x}_{iv}(0) &= \mathbf{0}, \\ \mathbf{c}_m &= \mathbf{N}_m^{\mathrm{T}} \mathbf{x}_r + \mathbf{c}_{0,m} \end{aligned} \tag{7}$$

Models (4) and (7) can be expressed in terms of the standard state-space equation form (1). Table 1 summarizes the two representations in the standard state-space form. Table 1 shows that the measurement function $\mathbf{g}(\cdot)$ is a linear function of the states.

Table 1. State space representations of batch reaction systems

Standard form	Model (4) Concentration domain	Model (7) Extent domain
$\mathbf{x}$	$\mathbf{c}$	$\mathbf{x}_r$
$\mathbf{y}$	$\mathbf{c}_m$	$\mathbf{c}_m$
$\mathbf{f}$	$\mathbf{N}^{\mathrm{T}}\mathbf{r}$	$\mathbf{r}$
$\mathbf{g}$	$\left[\mathbf{I}_m \;\; \mathbf{0}\right] \mathbf{c}$	$\mathbf{N}_m^{\mathrm{T}}\mathbf{x}_r + \mathbf{c}_{0,m}$
Number of states	S	R

3. IDENTIFIABILITY ANALYSIS OF REACTION NETWORKS

In this section, we demonstrate how the use of reaction variants or extents allows us to readily generate the characteristic set that significantly reduces the effort in deciding parameter identifiability in reaction systems.

In contrast to the conventional representation of reaction systems in the concentration domain, the same system is represented by a lower-dimensional subspace ($R-$dimensional) in the extent domain. Hence, the representation in terms of the extent reduces the complexity of the reaction system models when the differential-algebraic approach is applied to the model in the extent domain. Further, in the DAISY-based techniques, the Ritt's pseudo-division algorithm has to be applied to obtain the characteristic set of equations using symbolic computational techniques. In the extend-based approach, the characteristic equations can be obtained by solving a set of linear equations. This indeed reduces the computational efforts. These observations can be generalized as follows.

Proposition 7. (Varghese et al., 2018) Consider a reaction system with S species and R independent reactions. The reaction system can be represented by either the model (4) or the model (7). Let $\mathbf{y} = \mathbf{c}_m$ be the m-dimensional vector of the measured concentrations with $m \geq R$. Let $\mathbf{N}$ and $\mathbf{c}$ be partitioned as: $\mathbf{N} = [\mathbf{N}_m \;\; \mathbf{N}_u]$ and $\mathbf{c}^{\mathrm{T}} = [\mathbf{c}_m^{\mathrm{T}} \;\; \mathbf{c}_u^{\mathrm{T}}]$. If $\operatorname{rank}(\mathbf{N}_m) = R$, then the characteristic set can be given by the following equations:

$$(\mathbf{N}_m^T)^{\dagger} \dot{\mathbf{y}} = \mathbf{r}(\mathbf{y}, \boldsymbol{\theta}, \mathbf{c}_0) \tag{8}$$

Proposition 7 is used to determine the characteristic set without performing the Ritt's algorithm. for parameter identifiability in reaction systems. The result has been extended to systems with dependent reactions Varghese et al. (2018).

3.1 Mass action kinetics

A reaction is said to follow mass action kinetics if the rate law can be expressed as follows:

$$r = \prod_{i=1}^{S} k_f c_i^{|\nu_i|} - \prod_{j=1}^{S} k_b c_j^{|\nu_j|} \tag{9}$$

where ν_i and ν_j are the stoichiometric coefficients of the ith reactant and jth product in the reaction, and k_f and k_b are the forward and backward reaction constants. Consider a reaction system with S species and R independent reactions occurring in a constant volume batch reactor (for ease of exposition) which can be described by S differential equations:

$$\dot{\mathbf{c}}(t) = \mathbf{N}^{\mathrm{T}} \mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}), \quad \mathbf{c}(0) = \mathbf{c}_0 \tag{10}$$

where the rate laws follow mass action kinetics and $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of parameters to be determined. It is assumed that R species are measured such that conditions of proposition 7 are satisfied. From Proposition 7, we have

$$\dot{\mathbf{c}}_m = \mathbf{f}(\boldsymbol{\theta}, \mathbf{c}_m) \tag{11}$$

where $\mathbf{c}_m$ is the vector of measured concentrations. Clearly, the right hand side is a polynomial function of $\mathbf{c}_m$ and linear function of parameters θ. We collect the different monomials of the form $c_1^{i_1} c_2^{i_2} \dots c_{S_m}^{i_{S_m}}$ and examine the first row of $\mathbf{f}$ as follows:

$$\mathbf{f}(\boldsymbol{\theta}, \mathbf{c}_m)_1 = \left[(c_1^{i_1} \dots c_{S_m}^{i_{S_m}}) \dots (c_1^{l_1} \dots c_{S_m}^{l_{S_m}}) \right] \begin{bmatrix} \sum \alpha_{mi} \theta_i \\ \vdots \\ \sum \alpha_{li} \theta_i \end{bmatrix} \tag{12}$$

Let A_1 be defined as follows:

$$\begin{bmatrix} \sum \alpha_{mi} \theta_i \\ \vdots \\ \sum \alpha_{li} \theta_i \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha_{i1} & \alpha_{i2} & \dots & \alpha_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{l1} & \alpha_{m2} & \dots & \alpha_{lp} \end{bmatrix}}_{A_1} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} = A_1 \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \tag{13}$$

For every row of $\mathbf{f}_i$, $i = 1, \dots, R$, we collect the matrices A_i and form a matrix

$$\mathbf{h} = \begin{bmatrix} A_1 \\ \vdots \\ A_R \end{bmatrix}$$

Let us assume that in addition to measuring the concentrations $c_1, \dots, c_{S_m}$ at times $t_1, t_2, \dots, t_N$, the derivatives are also available at these times. Hence, we can form the following set of equations:

$$\begin{bmatrix} \dot{\mathbf{c}}(t_1) \\ \dot{\mathbf{c}}(t_2) \\ \vdots \\ \dot{\mathbf{c}}(t_N) \end{bmatrix} = \begin{bmatrix} \vdots & & \vdots \\ (c_1^{i_1} c_2^{i_2} \dots c_{S_m}^{i_{S_m}})(t_1) & \dots & (c_1^{l_1} c_2^{l_2} \dots c_{S_m}^{l_{S_m}})(t_1) \\ \vdots & & \vdots \\ (c_1^{i_1} c_2^{i_2} \dots c_{S_m}^{i_{S_m}})(t_2) & \dots & (c_1^{l_1} c_2^{l_2} \dots c_{S_m}^{l_{S_m}})(t_2) \\ \vdots & & \vdots \\ (c_1^{i_1} c_2^{i_2} \dots c_{S_m}^{i_{S_m}})(t_N) & \dots & (c_1^{l_1} c_2^{l_2} \dots c_{S_m}^{l_{S_m}})(t_N) \end{bmatrix} \mathbf{h}\boldsymbol{\theta} \tag{14}$$

Proposition 8. In reaction systems with mass action kinetics, the parameters are generically identifiable if matrix $\mathbf{h}$ in (14) is full rank, i.e., rank $(\mathbf{h}) = p$.

Proof In the interest of brevity, only an outline is presented. Clearly, Eq. (14) is a set of linear equations in $\boldsymbol{\theta}$ and over-determined, i.e., there are more equations than unknowns. The data matrix, i.e., the matrix of concentrations (or rather monomials) is of full column rank with high probability. Hence, if $\mathbf{h}$ is full rank (i.e., p), there exists a unique solution to the above equations, and parameters are identifiable. □

Consider an isothermal and constant density reaction system having $\mathbf{S}$ species and $\mathbf{R}$ independent reactions. Let this reaction network consist only of uni- or bimolecular reactions alone, i.e., reactions with a maximum of two reactants and two products. The general form of the ith reaction in such a network is given by:

$$R_i : R1_i + R2_i \leftrightarrow P1_i + P2_i$$

The rate of this ith reaction can be expressed in mass action kinetics as

$$r_i = k_{f,i} c_{R1,i}^{\alpha_i} c_{R2,i}^{\beta_i} - k_{b,i} c_{P1,i}^{\gamma_i} c_{P2,i}^{\theta_i} \tag{15}$$

Given this class of systems, we can show that the above system is identifiable if R measurements satisfying a rank condition are chosen.

Theorem 9. Consider a uni-/bi-molecular reaction system with S species and R independent reactions. Let $\mathbf{y} = \mathbf{c}_m$ be the R-dimensional vector of the measured concentrations. Let $\mathbf{N}$ and $\mathbf{c}$ be partitioned as: $\mathbf{N} = [\mathbf{N}_m \ \mathbf{N}_u]$ and $\mathbf{c} = [\mathbf{c}_m \ \mathbf{c}_u]^T$. If rank $(\mathbf{N}_m) = R$, $\mathbf{r}(\mathbf{y}; \mathbf{p}; \mathbf{c}_0) = 0$, and each reaction rate follows mass-action kinetics, then, the reaction system is globally identifiable. Alternatively, this result can be interpreted in the following manner. To guarantee parameter identifiability of uni-/bi-molecular reaction networks with mass-action kinetics, the minimum number of measurements for identifiability are R.

Proof In the interest of brevity, only an outline is presented. From Proposition 6, we have

$$\begin{aligned} \dot{x}_{r,i} &= r_i(\mathbf{x}_r, \boldsymbol{\theta}) \\ (\mathbf{N}_m^{-T})_i \dot{\mathbf{y}} &= \mathbf{r}(\mathbf{N}_m^{-T}(\mathbf{y} - \mathbf{c}_{m,0}), \boldsymbol{\theta}) \\ r_i &= k_{f,i} c_i^{\alpha_i} c_{R2,i}^{\beta_i} - k_{b,i} c_{P1,i}^{\gamma_i} c_{P2,i}^{\theta_i} \end{aligned} \tag{16}$$

We note that as all reactions are independent, k_{f_i} and k_{b_i} will only appear in the characteristic differential equation for the ith reaction. Hence, the identifiability of each set of k_{f_i} and k_{b_i} can be judged on the basis of whether or not the coefficients of the variables of a single equation in the characteristic set are alone globally identifiable or not. In the event all the species in a reaction are measured, we directly know that the coefficients are k_{f_i} and k_{b_i}, whose the Grobner basis will also simply be k_{f_i} and k_{b_i}, making them globally(uniquely) identifiable. A situation where all species in a reaction are unmeasured will not arise as this will violate our initial condition for measurement that rank$(\mathbf{N}_m)$=R. As the number of unmeasured species in any reaction in the network might vary, we still need to consider the following remaining cases:
(a) Three of the four species in a reaction are measured
(b) Two of the four species in a reaction are measured
(c) Only one species in a reaction is measured.
By analyzing each of these cases and following algebraic manipulations and simplifications, it is shown that the parameters in the ith reaction. Hence, the entire system is identifiable if R independent measurements are chosen. □

4. CONCLUSIONS

In this work, we showed that identifiability in reaction systems following mass action kinetics can be ascertained by examining the rank of a matrix thus obviating the need to determine the characteristic set using the differential algebraic methods. Furthermore, in order to guarantee parameter identifiability of uni-/bi-molecular reaction networks with mass-action kinetics, the minimum number of concentration measurements required for identifiability is R. This result can be used for designing experiments for kinetic model identification.

REFERENCES

Amrhein, M., Bhatt, N., Srinivasan, B., and Bonvin, D. (2010). Extents of reaction and flow for homogeneous reaction systems with inlet and outlet streams. *AIChE Journal*, 56(11), 2873–2886.

Bellu, G., Saccomani, M.P., Audoly, S., and D'Angiò, L. (2007). Daisy: A new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, 88(1), 52–61.

Bhatt, N.P. (2011). *Extents of Reaction and Mass Transfer in the Analysis of Chemical Reaction Systems, Doctoral Thesis No. 5028*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland.

Chis, O.T., Banga, J.R., and Balsa-Canto, E. (2011). Structural identifiability of systems biology models: a critical comparison of methods. *PloS ONE*, 6(11), e27755.

Jain, R., Narasimhan, S., and Bhatt, N.P. (2019). A priori parameter identifiability in models with non-rational functions. *Automatica*, 109, 108513.

Ljung, L. and Glad, T. (1994). On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2), 265–276.

Ritt, J.F. (1950). *Differential algebra*, volume 33. American Mathematical Soc.

Saccomani, M.P., Audoly, S., Bellu, G., and D'Angiò, L. (2010). Examples of testing global identifiability of biological and biomedical models with daisy software. *Computers in Biology and Medicine*, 40, 402–407.

Saccomani, M.P., Audoly, S., and D'Angiò, L. (2003). Parameter identifiability of nonlinear systems: the role of initial conditions. *Automatica*, 39(4), 619–632.

van Riel, N.A. (2006). Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Briefings in bioinformatics*, 7(4), 364–374.

Varghese, A., Narasimhan, S., and Bhatt, N. (2018). A priori parameter identifiability in complex reaction networks. *IFAC-PapersOnLine*, 51(15), 760–765. 18th IFAC Symposium on System Identification SYSID 2018.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Risk averse optimization with tensor decompositions ⋆

Harbir Antil * **Sergey Dolgov** ** **Akwum Onwunta** ***

* *The Center for Mathematics and Artificial Intelligence (CMAI) and Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030, USA (e-mail: hantil@gmu.edu).*
** *Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK (e-mail: s.dolgov@bath.ac.uk)*
*** *Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015, USA (e-mail: ako221@lehigh.edu)*

Abstract: We develop a new algorithm named TTRISK to solve high-dimensional risk-averse optimization problems governed by differential equations (ODEs and/or PDEs) under uncertainty. As an example, we focus on the so-called Conditional Value at Risk (CVaR), but the approach is equally applicable to other coherent risk measures. Both the full and reduced space formulations are considered. The algorithm is based on low rank tensor approximations of random fields discretized using stochastic collocation. To avoid non-smoothness of the objective function underpinning the CVaR, we propose an adaptive strategy to select the width parameter of the smoothed CVaR to balance the smoothing and tensor approximation errors. Moreover, unbiased Monte Carlo CVaR estimate can be computed by using the smoothed CVaR as a control variate. To accelerate the computations, we introduce an efficient preconditioner for the KKT system in the full space formulation.The numerical experiments demonstrate that the proposed method enables accurate CVaR optimization constrained by large-scale discretized systems. In particular, the first example consists of an elliptic PDE with random coefficients as constraints. The second example is motivated by a realistic application to devise a lockdown plan for United Kingdom under COVID-19. The results indicate that the risk-averse framework is feasible with the tensor approximations under tens of random variables.
This is an extended abstract for a talk based on https://arxiv.org/abs/2111.05180

Keywords: risk measures, tensor decompositions, function approximations, reduced space, preconditioning
MSC: 93E20, 65D15, 15A69

The control or design optimization problems constrained by stochastic systems must produce controls or optimal designs which are resilient to the uncertainty due to stochasticity. To tackle this, risk-averse optimization frameworks targeting engineering applications were created. This talk will introduce a new algorithm TTRISK which uses a Tensor Train (TT) decomposition to solve risk averse optimization problems constrained by differential equations (ODEs and/or PDEs).

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a complete probability space. Let U, Y be real reflexive Banach spaces, and let Z be a real Banach space. Here Y denotes the deterministic state space, U is the space of optimization variables (control or designs etc.) and Z is the differential equation residual space. Let $U_{ad} \subseteq U$ be a closed convex subset and let $c : Y \times U_{ad} \times \Omega \to Z$ denote, e.g., a PDE in a weak form, then consider the equality constraint

$$c(y, u; \omega) = 0, \quad \text{in } Z, \quad \text{a.a. } \omega \in \Omega$$

where a.a. indicates "almost all" with respect to a probability measure $\mathbb{P}$.

For practical computations we make the *finite dimensional noise assumption* on the equality constraint (Kouri and Surowiec (2016)). More precisely, the constraint $c(y, u; \omega) = 0$ is represented by a finite random vector $\xi : \Omega \to \Xi$, where $\Xi := \xi(\Omega) \subset \mathbb{R}^d$ with $d \in \mathbb{N}$. This allows us to redefine the probability space to (Ξ, Σ, ρ), where $\Sigma = \xi(\mathcal{A})$ is the σ-algebra of regions, and $\rho(\xi)$ is the continuous probability density function such that $\mathbb{E}[X] = \int_\Xi X(\xi)\rho(\xi)d\xi$. The random variable $X(\xi)$ can be considered as a function of the random vector $\xi = (\xi^{(1)}, \ldots, \xi^{(d)})$, belonging to the Hilbert space $\mathcal{F} = \{X(\xi) : \|X\| < \infty\}$, equipped with the inner product $\langle X, Y\rangle = \int_\Xi X(\xi)Y(\xi)\rho(\xi)d\xi$ and the Euclidean norm $\|X\| = \sqrt{\langle X, X\rangle}$.

Now we consider optimization problems of the form

$$\min_{u \in U_{ad}} \mathcal{R}[\mathcal{J}(y, u; \xi)] + \alpha\mathcal{P}(u) \tag{1}$$

⋆ HA and AO are partially supported by NSF grants DMS-2110263, DMS-1913004 and the Air Force Office of Scientific Research under Award NO: FA9550-19-1-0036. SD is thankful for the support from Engineering and Physical Sciences Research Council (EPSRC) New Investigator Award EP/T031255/1 and New Horizons grant EP/V04771X/1.

subject to $c(y,u;\xi)=0$, where $u\in U_{ad}$ is the deterministic control and $y\in Y$ is the state, $\mathcal{P}$ is the cost of the control, $\alpha\geq 0$ is the regularization parameter, $\mathcal{J}$ is the uncertain variable objective function and $\mathcal{R}$ is the *risk-measure* functional which maps random variables to extended real numbers.

We assume that $\mathcal{R}$ is based on expectation, i.e.,

$$\mathcal{R}[X]=\inf_{t\in\mathcal{T}}\mathcal{R}_t[X],\quad \text{where}\quad \mathcal{R}_t[X]:=\mathbb{E}[f(X,t)],\quad (2)$$

$f:\mathbb{R}\times\mathbb{R}^N\to\mathbb{R}$ and $\mathcal{T}\subseteq\mathbb{R}^N$, with $N\in\mathbb{N}$, is a closed convex set. A typical example of a risk measure $\mathcal{R}$ is the conditional value-at-risk (CVaR_β), where f in (2) is given by

$$f(X,t)=t+(1-\beta)^{-1}(X-t)_+,\quad (3)$$

with $\mathcal{T}=\mathbb{R}$, $\beta\in(0,1)$ is the confidence level and $(x)_+=\max\{x,0\}$. CVaR_β is also known as expected shortfall. It's origin lies in financial mathematics (Rockafellar and Uryasev (2000)), but owing to Kouri and Surowiec (2016), it is now being widely used in engineering applications. Our work in particular, focuses on minimization problems (1) with $\mathcal{R}$ given by CVaR_β but it can be extended to other coherent risk measures, such as buffered probability of exceedence (BPOE), of type (2).

Notice that, since risk measures, such as CVaR_β, focus on the upper tail events, the traditional sampling techniques to solve these stochastic PDE-constrained optimization problems are often computationally expensive. More precisely, CVaR_β captures the cost associated with rare events, but it requires more samples in order to be accurately approximated, which leads to many differential equation solves. Moreover, the presence of the non-smooth function $(\cdot)_+$ in CVaR_β poses several challenges, including, nondifferentiable cost functional, wasted Monte Carlo samples outside of the support of $(\cdot)_+=\max\{\cdot,0\}$, or slowly converging polynomial and other function approximation methods.

To tackle nonsmoothness in CVaR_β, Kouri and Surowiec (2016) has proposed a smoothing of $(\cdot)_+$ which requires solving a sequence of smoothed optimization problems using Newton-based methods. The smoothing approach is aimed at approximating a non-differentiable function $(\cdot)_+$ in CVaR_β by a smooth function $g_\varepsilon:\mathbb{R}\to\mathbb{R}$, which depends on some $\varepsilon>0$. In particular, we consider the following C^∞-smoothing function

$$g_\varepsilon(x)=\varepsilon\log(1+\exp(x/\varepsilon)).\quad (4)$$

Thus, the optimization problem for smooth $\text{CVaR}_\beta^\varepsilon$ is given by

$$\begin{cases}\min\limits_{(u,t)\in U_{ad}\times\mathbb{R}}\mathcal{R}^\varepsilon_{t,\beta}[\mathcal{J}(y,u;\xi)]+\alpha\mathcal{P}(u)\\ \text{subject to}\\ c(y,u;\xi)=0,\quad \text{in } Z,\quad \text{a.a. } \xi\in\Xi,\end{cases}\quad (5)$$

where

$$\mathcal{R}^\varepsilon_{t,\beta}[\mathcal{J}(y,u;\xi)]:=t+\frac{1}{1-\beta}\mathbb{E}[g_\varepsilon(\mathcal{J}(y,u;\xi)-t)].\quad (6)$$

We consider two formulations of (5). The first one is the *reduced-space* approach where we remove the equality constraint $c(y,u;\omega)=0$ via a control to solution map $u\mapsto y$. The second case is the full space approach, where we directly tackle the full problem (5) using the Lagrangian formulation. The latter formulation appears to be new in the context of risk-averse optimization. Numerical experiments demonstrate that the full formulation converges more reliably for extreme parameters, e.g. large β and small α.

To introduce the reduced-space formulation, we assume that $c(y,u;\xi)=0$ is uniquely solvable, i.e., for each $u\in U_{ad}$ there exists a unique solution mapping $y(u;\cdot):\Xi\to Y$ for almost all $\xi\in\Xi$. Moreover, we can approximate the exact expectation in (6) by a quadrature with some N points, $\mathbb{E}_N[f]\approx\mathbb{E}[f]$. The resulting optimization problem (5) only depends on u and is given by

$$\min_{(u,t)\in\mathcal{U}_{ad}\times\mathbb{R}}\left\{\mathfrak{J}_N(u,t):=\mathcal{R}^\varepsilon_{t,\beta,N}[j(u;\xi)]+\alpha\mathcal{P}(u)\right\},\quad (7)$$

where $j(u;\xi):=\mathcal{J}(y(u;\xi),u;\xi)$, and

$$\mathcal{R}^\varepsilon_{t,\beta,N}[j(u;\xi)]:=t+\frac{1}{(1-\beta)}\mathbb{E}_N[g_\varepsilon(j(u;\xi)-t)].$$

Computing first and second derivatives of $\mathfrak{J}_N(u,t)$, which exist in the classical sense for $\varepsilon>0$, we can formulate the Newton method.

In the full-space formulation, we introduce a Lagrange multiplier $p\in Z^*$, and a Lagrangian

$$\begin{aligned}\mathcal{L}_N(y,u,p,t)=t+(1-\beta)^{-1}\mathbb{E}_N\left[g_\varepsilon(\mathcal{J}(y,u,\xi)-t)\right]\\ +\alpha\mathcal{P}(u)+\mathbb{E}_N\langle p,c(y,u,\xi)\rangle,\end{aligned}\quad (8)$$

which is again differentiable for $\varepsilon>0$, so we can find a KKT point using the Newton method.

The dimension d of the random vector ξ can be arbitrarily high. For instance, ξ may be a tuple of tens of model tuning parameters, or it can be a vector of coefficients of a Karhunen-Loeve (KL) approximation of an infinite-dimensional continuous random field. In this case expectations as in (6) become high-dimensional integrals. Instead of a direct Monte Carlo average (which may converge too slowly), we introduce a high-order quadrature rule (e.g. Gauss-Legendre) with $n_\xi\in\mathbb{N}$ points in each of the components $\xi^{(1)},\dots,\xi^{(d)}$ independently. For this, we assume that each $\xi^{(k)}$ has a probability density function $\rho^{(k)}(\xi^{(k)})$, and that the space of functions $f(\xi)\in\mathcal{F}$ is isomorphic to a Cartesian product of spaces of univariate functions, $\mathcal{F}=\mathcal{F}^{(1)}\otimes\cdots\otimes\mathcal{F}^{(d)}$, where $\mathcal{F}^{(k)}=\{f^{(k)}(\xi^{(k)}):\|f^{(k)}\|<\infty\}$, $\|f^{(k)}\|=\sqrt{\langle f^{(k)},f^{(k)}\rangle}$, $\langle f^{(k)},g^{(k)}\rangle=\int_{\mathbb{R}}f^{(k)}(\xi^{(k)})g^{(k)}(\xi^{(k)})\rho^{(k)}(\xi^{(k)})d\xi^{(k)}$, $k=1,\dots,d$. However, the exponential total number of quadrature points in all variables n_ξ^d becomes intractable even for moderate dimensions.

To tackle this curse of dimensionality, we build on tensor decomposition methods, which emerged in the past two decades (Hackbusch (2012)) as an efficient approximation of multi-index arrays, in particular when those contain expansion coefficients of high-dimensional functions (Gorodetsky et al. (2019)).

Definition. A square-integrable function $f(\xi)$ is said to be approximated by a *(functional) TT decomposition* (Gorodetsky et al. (2019)) with a relative approximation error ϵ if there exist univariate functions $F^{(k)}(\cdot):\xi^{(k)}\in\mathbb{R}\to\mathbb{R}^{r_{k-1}\times r_k}$, $k=1,\dots,d$, such that

$$\tilde{f}(\xi) := \sum_{s_0,\dots,s_d=1}^{r_0,\dots,r_d} F^{(1)}_{s_0,s_1}(\xi^{(1)}) F^{(2)}_{s_1,s_2}(\xi^{(2)}) \cdots F^{(d)}_{s_{d-1},s_d}(\xi^{(d)}), \tag{9}$$

where the subscripts s_{k-1}, s_k denote elements of a matrix, and $\|f - \tilde{f}\| = \epsilon \|f\|$. The factors $F^{(k)}$ are called *TT cores*, and their image dimensions $r_0, \dots, r_d \in \mathbb{N}$ are called *TT ranks*.

Without loss of generality we can let $r_0 = r_d = 1$, but the other TT ranks $r_1, \dots, r_{d-1}$ can vary depending on the approximation error. The simplest example is a bi-variate truncated Fourier series $\tilde{f}(\xi^{(1)}, \xi^{(2)}) = \sum_{s=-r}^{r} f_s(\xi^{(1)}) \exp(\mathrm{i} s \xi^{(2)})$.

From (9), we notice that the expectation of $\tilde{f}$ factorizes into univariate integrations,

$$\mathbb{E}_N[f] = \mathbb{E}[\tilde{f}] = \sum_{s_0,\dots,s_d=1}^{r_0,\dots,r_d} (\mathbb{E}F^{(1)}_{s_0,s_1})(\mathbb{E}F^{(2)}_{s_1,s_2}) \cdots (\mathbb{E}F^{(d)}_{s_{d-1},s_d}).$$

For practical computations with (9) we introduce univariate bases $\{\ell_i(\xi^{(k)})\}_{i=1}^{n_\xi}$, and the multivariate basis constructed from a Cartesian product,

$$L_{i_1,\dots,i_d}(\xi) := \ell_{i_1}(\xi^{(1)}) \cdots \ell_{i_d}(\xi^{(d)}).$$

Now we can collect the expansion coefficients of $\tilde{f}$ into a tensor $\mathbf{F} \in \mathbb{R}^{n_\xi \times \cdots \times n_\xi}$,

$$\tilde{f}(\xi) = \sum_{i_1,\dots,i_d=1}^{n_\xi} \mathbf{F}(i_1, \dots, i_d) L_{i_1,\dots,i_d}(\xi). \tag{10}$$

Similarly, TT cores in (9) can be written using three-dimensional tensors $\mathbf{F}^{(k)} \in \mathbb{R}^{r_{k-1} \times n_\xi \times r_k}$,

$$F^{(k)}_{s_{k-1},s_k}(\xi^{(k)}) = \sum_{i=1}^{n_\xi} \mathbf{F}^{(k)}(s_{k-1}, i, s_k) \ell_i(\xi^{(k)}), \tag{11}$$

$k = 1, \dots, d$. The original (discrete) TT decomposition (Oseledets and Tyrtyshnikov (2010)) was introduced to decompose a high-dimensional tensor like $\mathbf{F}$ into a product of smaller tensors like $\mathbf{F}^{(k)}$. Note that $\mathbf{F}$ contains n_ξ^d elements, whereas storing $\mathbf{F}^{(1)}, \dots, \mathbf{F}^{(d)}$ needs only $\sum_k r_{k-1} n_\xi r_k$ elements. For brevity we can define the maximal TT rank $r := \max_k r_k$, which gives us a linear storage complexity of the TT decomposition, $\mathcal{O}(d n_\xi r^2)$.

In practice, TT cores are computed by solving an approximation or optimization problem in an alternating direction-type iteration. For example, the *TT-Cross* method by Oseledets and Tyrtyshnikov (2010) can approximate potentially any function $\tilde{f}(\xi) \approx f(\xi)$, using $\mathcal{O}(d n_\xi r^2)$ samples from the function $f(\xi)$ and $\mathcal{O}(d n_\xi r^3)$ further floating point operations. Similarly, linear algebra on functions can be recast to linear algebra on their TT cores with a linear complexity in the dimension.

However, irregular functions, such as the $(\cdot)_+$ function in (3), may lack an efficient TT decomposition. This is another reason for switching to the smoothed CVaR formulation, since smooth functions do admit convergent TT approximation. Nevertheless, this convergence can still be slow, and for practically feasible smoothness parameters, the solution of the smoothed problem can be biased. To obtain an unbiased, asymptotically exact solution, we propose a version of Multilevel Monte Carlo methods (Giles (2015)), namely, we use a smoothed solution as a control variate.

The TTRISK algorithm is a combination of the Newton method for (7) or (8) with TT approximations (9) for the computation of expectations (with optional Monte Carlo correction), and an adaptive selection of ε which is driven to zero geometrically as the iteration converges.

The numerical experiments demonstrate that the stochastic risk-averse control problem can be solved with a cost that depends at most polynomially on the dimension. This allows us to solve realistic risk-averse PDE and ODE control problems with up to 20 random variables.

REFERENCES

Giles, M.B. (2015). Multilevel Monte Carlo methods. *Acta Numer.*, 24, 259–328.

Gorodetsky, A., Karaman, S., and Marzouk, Y. (2019). A continuous analogue of the tensor-train decomposition. *Comput. Methods Appl. Mech. Engrg.*, 347, 59–84. doi:10.1016/j.cma.2018.12.015. URL `https://doi.org/10.1016/j.cma.2018.12.015`.

Hackbusch, W. (2012). *Tensor Spaces And Numerical Tensor Calculus*. Springer–Verlag, Berlin.

Kouri, D.P. and Surowiec, T.M. (2016). Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM J. Optim.*, 26(1), 365–396. doi:10.1137/140954556. URL `https://doi.org/10.1137/140954556`.

Oseledets, I.V. and Tyrtyshnikov, E.E. (2010). TT-cross approximation for multidimensional arrays. *Linear Algebra Appl.*, 432(1), 70–88. doi:10.1016/j.laa.2009.07.024.

Rockafellar, R.T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk*, 2, 21 – 41.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A note on efficient and reliable prediction-based control in the Koopman framework ⋆

Manuel Schaller *, **Karl Worthmann** *, **Friedrich Philipp** *, **Sebastian Peitz** **, **and Feliks Nüske** **,***,

* *Technische Universität Ilmemau, Institute of Mathematics, Optimization-based Control group, Germany (e-mail: {friedrich.philipp, manuel.schaller, karl.worthmann}@tu-ilmenau.de).*
** *Paderborn University, Department of Computer Science, Data Science for Engineering, Germany (e-mail: sebastian.peitz@upb.de).*
*** *Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany (e-mail: nueske@mpi-magdeburg.mpg.de)*

Abstract: Extended Dynamic Mode Decomposition, embedded in the Koopman framework, is a widely-applied technique to predict the evolution of an observable along the flow of a dynamical (control) system. However, despite its popularity, the error analysis for control systems is still fragmentary. Here, we provide a complete and rigorous analysis of the approximation error for control systems. To this end, the approximation error is split up according to its two sources of error: the finite dictionary size (projection) and the finite amount of i.i.d. data used to generate the surrogate model (estimation). Then, invoking—among others—finite-elements techniques and the Chebyshev inequality, probabilistic error bounds are derived. Finally, we demonstrate the applicability of the novel error bounds in optimal control with state and control constraints.

Keywords: data-based control, eDMD, error bound, finite-data, Koopman, state constraints

While optimal and predictive control based on models derived from first principles are nowadays well established, data-driven controller design is becoming increasingly popular. A particularly successful technique to construct data-driven surrogate models is based on the extended Dynamic Mode Decomposition (eDMD), see Mezić (2005); Rowley et al. (2009) and the recent survey Brunton et al. (2021). A key advantage is the potential embedding of eDMD in the Koopman framework (Koopman, 1931) such that it has a sound theretical foundation.

This extended abstract, which is mainly a summary of the major findings presented in full detail in Schaller et al. (2022), provides a complete error analysis for the bilinear surrogate models resulting for nonlinear control-affine systems. To this end, the (overall) approximation error is split up into its two sources, i.e., projection and estimation. The projection error results from using a dictionary $D(\mathcal{L})$ consisting of only finitely many entries—the so-called observables $\psi_1, \dots, \psi_N$. The estimation error results from only taking finitely many data points $x_1, \dots, x_m$ in account while constructing the surrogate model on the subspace $\mathbb{V} = \operatorname{span}\{\{\psi_j\}_{j=1}^N\}$. Here, the dictionary includes linear finite elements on a triangulation of the domain of interest. In conclusion and to the best of the authors' knowledge, this is the first rigorous *finite-data* error estimate for the eDMD-based prediction for nonlinear control systems taking into account both sources of errors, i.e., the projection *and* the estimation error.

⋆ F. Philipp was funded by the Carl Zeiss Foundation within the project *DeepTurb—Deep Learning in und von Turbulenz*. K. Worthmann gratefully acknowledges funding by the German Research Foundation (DFG; grant WO 2056/6-1, project number 406141926).

1. BILINEAR SURROGATE MODEL

The Koopman framework provides the theoretical foundation for data-driven approximation techniques like eDMD, see (Mauroy et al., 2020, Chapters 1 and 8): Using the Koopman operator semi-group $(\mathcal{K}^t)_{t\geq 0}$ or—equivalently—the Koopman generator $\mathcal{L}$, so-called observables φ (real- or complex-valued L^2-functions of the state, e.g., representing a coordinate function, a state constraint or another quantity of interest) can be propagated forward in time via

$$\mathcal{K}^t\varphi = \mathcal{K}^0\varphi + \mathcal{L}\int_0^t \mathcal{K}^s\varphi \,\mathrm{d}s. \tag{1}$$

The propagated observable $\mathcal{K}^t\varphi$ can be evaluated for some state x_0 instead of first calculating the solution $x(t;x_0)$ of the underlying Ordinary Differential Equation (ODE) and then evaluating the observable as depicted in Figure 1.

The convergence of $(\tilde{\mathcal{K}}^t)_{t\geq 0}$, i.e., the eDMD-based surrogate model of an autonomous dynamical system, to the Koopman semi-group $(\mathcal{K}^t)_{t\geq 0}$ in the infinite-data limit, i.e., for N and m tending to infinity, was shown in Korda and Mezić (2018b). However, error estimates for the estimation step explicitly depending on the dictionary size N and the amount of data points m for the ODE case were only recently established in (Zhang and Zuazua, 2021) and (Nüske et al., 2021) assuming identically and independently distributed (i.i.d.) data. While the former reference

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

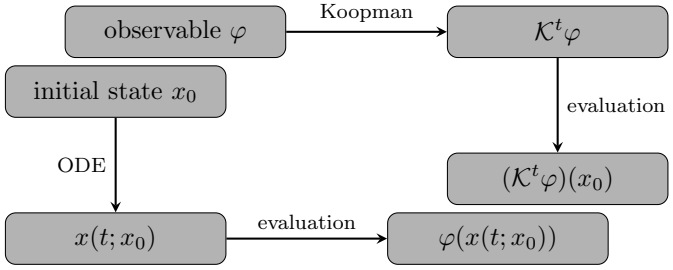


Fig. 1. Schematic sketch of the Koopman framework.

also covers the projection error in a rigorous manner, the latter one extends the analysis of the estimation error to stochastic differential equations and ergodic sampling.

We consider dynamics governed by the nonlinear control-affine differential equation

$$\dot{x}(t) = g_0(x(t)) + \sum_{i=1}^{n_c} g_i(x(t))u_i(t) \tag{2}$$

with initial condition $x(0) = x_0$ and locally Lipschitz-continuous vector fields $g_0, g_1, \ldots, g_{n_c} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$. Further, we impose the control constraints $u(t) \in \mathbb{U}$ for some compact, convex, and nonempty set $\mathbb{U} \subset \mathbb{R}^{n_c}$ and define the set of admissible control functions by

$$\mathcal{U}_T(x_0) \triangleq \left\{ u : [0,T] \to \mathbb{R}^{n_c} \;\middle|\; \begin{array}{l} u \text{ measurable} \\ \exists!\, x(\,\cdot\,;x_0,u) \\ u(t) \in \mathbb{U}, t \in [0,T] \end{array} \right\}, \tag{3}$$

where $x(t;x_0,u)$ denotes the unique solution at time $t \geq 0$. In the following, we assume that $\mathcal{U}_T(x_0)$ is nonempty whenever the set of initial values $\mathbb{X} \subset \mathbb{R}^{n_x}$ is suitably chosen.

In order to predict control systems by means of the Koopman framework, Proctor et al. (2016) as well as Korda and Mezić (2018a) proposed to augment the state by the control variable and, then, to generate a linear surrogate model depending on the extended state by using eDMD, see (Mauroy et al., 2020, Chapter 1) for details. In this work, however, we use the bilinear approach, exploiting the control-affine structure of (2) as suggested, e.g., in Williams et al. (2016); Surana (2016), see also (Mauroy et al., 2020, Section 4), for which estimation error bounds were first derived in Nüske et al. (2021). The advantages of this approach are twofold. First, one can observe a superior performance when considering bilinear systems where the control is coupled to the state. Second, as the state dimension is not augmented, the data-requirements are less demanding.

We briefly describe the bilinear approach for surrogate modeling of control systems. In this case, for a given control $u \in L^\infty(0,T;\mathbb{U})$, the generator inherits the control-affine structure, i.e.,

$$\mathcal{L}^u(t) = \mathcal{L}^0 + \sum_{i=1}^{n_c} u_i(t)\left(\mathcal{L}^{e_i} - \mathcal{L}^0\right), \tag{4}$$

with $\mathcal{L}^{e_i}$, $i = 0, \ldots, n_c$, being the Koopman generator for the autonomous system with constant control $\bar{u} = e_i$, where $e_0 := 0$ and e_j denotes the j-th standard basis vector, see, e.g., Peitz et al. (2020). Thus, the time evolution of an observable function $\varphi \in L^2(\mathbb{X})$ along the flow of the control system (2) can be predicted via the bilinear system

$$\dot{z}(t;u) = \mathcal{L}^u(t)z(t;u), \qquad z(0;u) = \varphi. \tag{5}$$

This propagated observable function can, then, be evaluated for an initial state $x_0 \in \mathbb{X}$ to evaluate the observable along a particular trajectory, i.e.,

$$z(t;u)(x_0) = \varphi(x(t;x_0,u)),$$

cf. Figure 1 for an illustration.

Each of the individual generators $\mathcal{L}^{e_i}$, $i \in \{0, \ldots, n_c\}$ can be approximated by means of eDMD. The orthogonal projection onto $\mathbb{V}$ and the Galerkin projection of the Koopman generator are denoted by $P_\mathbb{V}$ and $\mathcal{L}^{e_i}_\mathbb{V} := P_\mathbb{V}\mathcal{L}^{e_i}_{|\mathbb{V}}$, respectively. Along the lines of Klus et al. (2020), we have the representation $\mathcal{L}^{e_i}_\mathbb{V} = C^{-1}A$ with $C, A \in \mathbb{R}^{N\times N}$ given by

$$C_{i,j} = \langle \psi_i, \psi_j \rangle_{L^2(\mathbb{X})} \quad \text{and} \quad A_{i,j} = \langle \psi_i, \mathcal{L}^{e_i}\psi_j \rangle_{L^2(\mathbb{X})}.$$

Consider i.i.d. data points $x_1, \ldots, x_m \in \mathbb{X}$ and the matrices

$$\Psi(X) := \left(\begin{pmatrix} \psi_1(x_1) \\ \vdots \\ \psi_N(x_1) \end{pmatrix} \middle| \ldots \middle| \begin{pmatrix} \psi_1(x_m) \\ \vdots \\ \psi_N(x_m) \end{pmatrix} \right)$$

$$\mathcal{L}^{e_i}\Psi(X) := \left(\begin{pmatrix} (\mathcal{L}^{e_i}\psi_1)(x_1) \\ \vdots \\ (\mathcal{L}^{e_i}\psi_N)(x_1) \end{pmatrix} \middle| \ldots \middle| \begin{pmatrix} (\mathcal{L}^{e_i}\psi_1)(x_m) \\ \vdots \\ (\mathcal{L}^{e_i}\psi_N)(x_m) \end{pmatrix} \right).$$

Then, defining $\tilde{C}_m, \tilde{A}_m \in \mathbb{R}^{N\times N}$ by

$$\tilde{C}_m = \tfrac{1}{m}\Psi(X)\Psi(X)^\top \quad \text{and} \quad \tilde{A}_m = \tfrac{1}{m}\Psi(X)\mathcal{L}^{e_i}\Psi(X)^\top,$$

an empirical, i.e., purely data-based estimator for the Galerkin projection $\mathcal{L}_\mathbb{V}$ is given by $\tilde{\mathcal{L}}_m = \tilde{C}_m^{-1}\tilde{A}_m$. Hence, the projection of (4) onto the finite dictionary $\mathbb{V}$ is given by

$$\mathcal{L}^u_\mathbb{V}(t) := \mathcal{L}^0_\mathbb{V} + \sum_{i=1}^{n_c} u_i(t)\big(\mathcal{L}^{e_i}_\mathbb{V} - \mathcal{L}^0_\mathbb{V}\big). \tag{6}$$

In conclusion, the propagation of observable functions $\varphi \in L^2(\mathbb{X})$ projected onto the dictionary $\mathbb{V}$ is given by

$$\dot{z}_\mathbb{V}(t;u) = \mathcal{L}^u_\mathbb{V}(t)z_\mathbb{V}(t;u), \qquad z(0;u) = \mathbb{P}_\mathbb{V}\varphi. \tag{7}$$

The corresponding approximation by means of eDMD using m data points is defined analogously via

$$\tilde{\mathcal{L}}^u_m(t) := \tilde{\mathcal{L}}^0_m + \sum_{i=1}^{n_c} u_i(t)\big(\tilde{\mathcal{L}}^{e_i}_m - \tilde{\mathcal{L}}^0_m\big), \tag{8}$$

where for $i = 1, \ldots, n_c$, $\tilde{\mathcal{L}}^{e_i}_m$ are eDMD-based approximations of $\mathcal{L}^{e_i}_\mathbb{V}$. The corresponding data-based surrogate prediction dynamics read

$$\dot{\tilde{z}}_m(t;u) = \tilde{\mathcal{L}}^u_m(t)\tilde{z}_m(t;u) \qquad \tilde{z}_m(0;u) = \mathbb{P}_\mathbb{V}\varphi. \tag{9}$$

2. UNIFORM FINITE-DATA ERROR BOUND

In this section, we present error bounds when using a dictionary $\mathbb{V}$ consisting of finite elements, i.e., $\psi_i, i = 1, \ldots, N$, are the usual piecewise linear hat functions defined on a regular uniform triangulation of the compact set $\mathbb{X}$ with meshsize $\Delta x > 0$ (e.g., the maximal incircle diameter of the cells) and nodes $\hat{x}_j \subset \mathbb{X}$, $j = 1, \ldots, N$ such that for $i, j \in \{1, \ldots, N\}$, $\psi_i(\hat{x}_j) = 1$ if $i = j$ and zero otherwise.

We now state the main theorem considering the approximation error.

Theorem 1. Let a probabilistic tolerance $\varepsilon > 0$ and confidence level $1 - \delta \in (0,1)$ be given. Then, for an observable $\varphi \in \mathcal{C}^2(\mathbb{X},\mathbb{R})$, $\mathbb{X} \subset \mathbb{R}^n$ compact and time $T > 0$, there exist

constants c_{dict}, c_{data} such that the probabilistic bound on the approximation error

$$\mathbb{P}\left(\|\varphi(x(t;x_0,u)) - \tilde{z}_m(t;u)(x_0)\| \leq \varepsilon\right) \geq 1-\delta$$

holds for all $x_0 \in \mathbb{X}$, $u \in \mathcal{U}_T(x_0)$ and all $t \in [0,T]$ provided that $\{x(t;x_0,u) : t \in [0,T]\} \subset \mathbb{X}$ for the bilinear surrogate model (9) if the mesh size Δx of the finite-elements observables and the number of i.i.d. data points satisfy the data-requirement conditions

$$\Delta x \leq c_{\text{dict}}\varepsilon, \qquad m \geq c_{\text{data}} \frac{1}{\varepsilon^2 \delta} \left(\tfrac{1}{\Delta x^d}\right)^2 \tag{10}$$

Sketch of the proof. The proof follows the same argumentation as (Schaller et al., 2022, Proof of Theorem 5). The main idea is sketched in Figure 2: One first fixes a (small enough) finite element mesh size Δx, and hence the dictionary, such that the projection error is small enough. This can be done using standard finite element estimates. Then, having fixed the dictionary size, using the Chebyshev inequality, one chooses a (high enough) number of data points such that also the estimation error on this dictionary is small enough. Due to the randomness in the data points, this can only be obtained with a given desired confidence, leading to the probabilistic estimate in the statement.

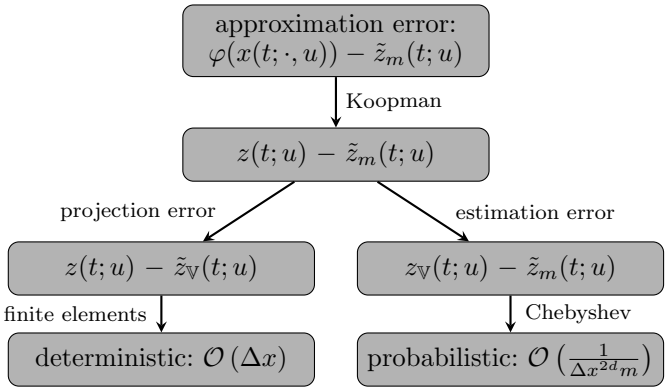


Fig. 2. Splitting of full error into projection and estimation error.

3. UNIFORM ERROR BOUNDS IN OPTIMAL CONTROL

We consider the Optimal Control Problem

$$\text{Minimize}_{u \in \mathcal{U}_T(x_0)} \int_0^T \ell(x(t;x_0,u)) + \|u(t)\|_2^2 \,\mathrm{d}t \quad \text{(OCP)}$$

subject to the initial condition $x(0) = x_0$, the control affine system dynamics (2), and the state constraints

$$h_j(x(t;x_0,u)) \leq 0 \qquad \forall\, j \in \{1,2,\dots,p\}. \tag{11}$$

The key challenge is to properly predict the performance index of (OCP) and satisfaction of the state constraints (11). Both quantities are evaluated along the state dynamcis (2), i.e., the stage cost ℓ and the constraint functions h_j, $j = 1,2,\dots,p$. Instead of propagating the state dynamics and then evaluating these *observables*, we employ the derived eDMD-based surrogate model to obtain the equality

$$(\mathcal{K}_u^t \varphi)(x_0) = \varphi(x(t;x_0,u)) \qquad \forall\, t \in [0,T] \tag{12}$$

for the observables $\varphi = h_j$, $j \in \{1,\dots,p\}$, and $\varphi = \ell$. Then, suitably applying Theorem 1 yields the error bounds, see (Schaller et al., 2022, Proposition 7).

Proposition 2. Suppose that ℓ, $h_i \in C^2(\mathbb{X},\mathbb{R})$, $i \in \{1,2,\dots,p\}$, hold. Then, for every probabilistic tolerance $\varepsilon > 0$ and every confidence level $1-\delta \in (0,1)$, every initial value $x_0 \in \mathbb{X}$ such that $x(t;x_0,u) \in \mathbb{X}$ for the solution of (2) contained in $\mathbb{X}$,

1) the probabilistic *performance bound*
$$\mathbb{P}\left(\|\ell(x(t;x_0,u)) - \tilde{\ell}_m(t;x_0,u)\| \leq \varepsilon\right) \geq 1-\delta$$
2) and the probabilistic state-constraint satisfaction, i.e.,
$$\mathbb{P}(h_i(x(t,x_0;u)) \leq 0) \geq 1-\delta,$$
is satisfied provided that $\tilde{h}_{i,m}(t;x_0,u) \leq -\varepsilon$ holds,

where $\tilde{\ell}_{1,m}$, $\tilde{h}_{i,m}$, $i \in \{1,2,\dots,p\}$, are predicted along the bilinear surrogate dynamics (9) with initial states $\tilde{\ell}_{1,m}(0;x_0,u) = P_{\mathbb{V}}\ell_1$, $\tilde{h}_{i,m}(0;x_0,u) = P_{\mathbb{V}}h_i$ if the requirements (10) for dictionary and data are met.

The result of Proposition 2 is two-fold. The first statement 1) yields a bound on the stage cost. This can be utilized to estimate the degree of suboptimality of the surrogate OCP's optimal control in view of the original problem. The second statement 2) allows us to deduce a chance constraint satisfaction of the original problem, provided we satisfy tightened constraints along the surrogate model.

4. SUMMARY AND OUTLOOK

We provided a complete and an rigorous analysis of the approximation error for eDMD-based prediction of control systems in the Koopman framework. To this end, we split the error into its two sources of error, a projection error vanishing in the dictionary size and an estimation error decreasing in the number of data points. Last, we applied the error bound to particular observables in optimal control, namely the stage cost and the constraint functions, to derive probabilistic error bounds on the performance index and the satisfaction of state constraints.

In future work, we want to apply the derived results, cf. Proposition 2, within data-driven predictive control. In particular, we aim at showing recursive feasibility and deriving suboptimality estimates in data-based predictive control.

REFERENCES

Brunton, S.L., Budišić, M., Kaiser, E., and Kutz, J.N. (2021). Modern Koopman theory for dynamical systems. Preprint available at: arXiv:2102.12086.

Klus, S., Nüske, F., Peitz, S., Niemann, J.H., Clementi, C., and Schütte, C. (2020). Data-driven approximation of the Koopman generator: Model reduction, system identification, and control. *Physica D*, 406, 132416.

Koopman, B.O. (1931). Hamiltonian Systems and Transformations in Hilbert Space. *Proc. National Academy of Sciences*, 17(5), 315–318.

Korda, M. and Mezić, I. (2018a). Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93, 149–160.

Korda, M. and Mezić, I. (2018b). On Convergence of Extended Dynamic Mode Decomposition to the Koopman Operator. *J. Nonlinear Science*, 28(2), 687–710.

Mauroy, A., Susuki, Y., and Mezić, I. (2020). *Koopman operator in systems and control*. Springer.

Mezić, I. (2005). Spectral Properties of Dynamical Systems, Model Reduction and Decompositions. *Nonlinear Dynamics*, 41, 309–325.

Nüske, F., Peitz, S., Philipp, F., Schaller, M., and Worthmann, K. (2021). Finite-data error bounds for Koopman-based prediction and control. Preprint available at: arXiv:2108.07102.

Peitz, S., Otto, S.E., and Rowley, C.W. (2020). Data-driven model predictive control using interpolated Koopman generators. *SIAM J. Applied Dynamical Systems*, 19(3), 2162–2193.

Proctor, J.L., Brunton, S.L., and Kutz, J.N. (2016). Dynamic mode decomposition with control. *SIAM J. Applied Dynamical Systems*, 15(1), 142–161.

Rowley, C.W., Mezić, I., Bagheri, S., Schlatter, P., and Henningson, D.S. (2009). Spectral analysis of nonlinear flows. *J. Fluid Mechanics*, 641, 115–127.

Schaller, M., Worthmann, K., Nüske, F., Peitz, S., and Philipp, F. (2022). Towards efficient and reliable prediction-based control using edmd. Preprint available at `https://doi.org/10.48550/arXiv.2202.09084`.

Surana, A. (2016). Koopman operator based observer synthesis for control-affine nonlinear systems. In *Proc. 55th IEEE Conf. Decision Control (CDC)*, 6492–6499.

Williams, M.O., Hemati, M.S., Dawson, S.T., Kevrekidis, I.G., and Rowley, C.W. (2016). Extending data-driven Koopman analysis to actuated systems. *IFAC-PapersOnLine*, 49(18), 704–709.

Zhang, C. and Zuazua, E. (2021). A quantitative analysis of Koopman operator methods for system identification and predictions. Preprint available at: hal-03278445.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Energy shaping, Interconnection and entropy assignment of boundary controlled irreversible port-Hamiltonioan systems: the heat equation

Y. Le Gorrec * **L. Mora Araque** ** **H. Ramirez** ***

* *FEMTO-ST AS2M, UBFC, ENSMM, 26 chemin de l'Epitaphe, F-25030 Besançon, France, (e-mail:legorrec@femto-st.fr)*
** *Dept. of Applied Mathematics,University of Waterloo, Waterloo, ON N2L 3G1 Canada, (e-mail: luis.mora@uwaterloo.ca)*
*** *C3E, Universidad Tecnica Federico Santa Maria, Av. Espana 1680, Valparaiso, Chile, (e-mail: hector.ramireze@usm.cl)*

Abstract: In this extended abstract we show, on the one dimensional (1D) heat equation example, how Boundary Controlled Irreversible Port Hamiltonian Systems (BC-IPHS) formulations and systems thermodynamic fundamental properties can be used for control design purposes.

Keywords: Irreversible port Hamiltonian systems, Boundary control systems, Heat equation.

1. INTRODUCTION

Boundary controlled port Hamiltonian formulations [4, 2] have shown to be very useful for the analysis and control of distributed parameter systems described by skew symmetric differential operators. For this class of systems, an efficient control design technique has been established using control by interconnection, structural invariants and energy shaping [5]. These formulations have been recently extended to irreversible systems [6], *i.e.* systems for which the thermal domain plays a central role, as it is the case for the heat equation or for reaction-diffusion systems. Boundary Controlled Irreversible Port Hamiltonian Systems (BC-IPHS) formulations allow to encompass the two laws of Thermodynamics, conservation of the total energy and irreversible entropy creation. In this extended abstract we show on the one dimensional (1D) heat equation example how these fundamental properties can be used for control design purpose.

2. BOUNDARY CONTROLLED IPHS FORMULATION OF THE HEAT EQUATION:

Alternatively to the classical formulation of the heat equation defined on a one dimensional spatial domain, with $\zeta \in [0, L]$,

$$c_v \partial_t T = \partial_\zeta \left(k \partial_\zeta T \right), \tag{1}$$

where $T = T(\zeta, t)$ is the temperature, c_v the heat capacitance of the medium and k the heat conduction coefficient, one can use the entropy density $s = s(\zeta, t)$ as state variable. Using the expression of the entropy flux $q_s = \frac{q}{T}$ with $q = -k\partial_\zeta T$ and Gibbs relation $du = Tds$, (1) can be written

$$\partial_t s = -\partial_\zeta q_s - \left(\frac{q_s}{T} \right) \partial_\zeta T, \tag{2}$$

that leads after integration by parts, according to [6], to the BC-IPHS

$$\begin{aligned} \partial_t s =& r_s \partial_\zeta e_s + \partial_\zeta \left(r_s e_s \right) \\ \mathbf{u} =& W_B e_s = \begin{bmatrix} -q_s|_L \\ -q_s|_0 \end{bmatrix}, \\ \mathbf{y} =& W_C e_s = \begin{bmatrix} T|_L \\ -T|_0 \end{bmatrix} \end{aligned} \tag{3}$$

with [1] $r_s = \gamma_s \{\mathcal{S}|\mathcal{H}\}$, $\gamma_s = \frac{k}{T^2}$, $e_s = \delta_s \mathcal{H} = T$ and

$$W_B = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad W_C = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

The total energy and total entropy are given by $\mathcal{H} = \int_0^L u(\zeta, t) \mathrm{d}\zeta$ and $\mathcal{S} = \int_0^L s(\zeta, t) \mathrm{d}\zeta$ respectively, such that $\delta_s \mathcal{H} = \partial_s u = T$ and $\delta_s \mathcal{S} = 1$. The BC-IPHS (3) expresses explicitly the first and second laws of Thermodynamics. Indeed, the term $r_s \partial_\zeta e_s = \gamma_s \{\mathcal{S}|\mathcal{H}\}^2 \geq 0$ describes the density of entropy produced by the heat flux and the term $\partial_\zeta \left(r_s e_s \right) = -\partial_\zeta q_s$ describes the entropy diffusion, so the total entropy and total energy balance are

$$\dot{\mathcal{S}} = \int_0^L \partial_t s \mathrm{d}\zeta = \underbrace{\int_0^L \gamma_s \{\mathcal{S}|\mathcal{H}\}^2 \mathrm{d}\zeta}_{\geq 0} + [1 \ -1] \, \mathbf{u} \tag{4}$$

and

$$\begin{aligned} \dot{\mathcal{H}} &= \int_0^L e_s \partial_t s \mathrm{d}\zeta = \int_0^L \left(e_s r_s \partial_\zeta \left(e_s \right) + \delta_s \mathcal{H} \partial_\zeta \left(r_s e_s \right) \right) \mathrm{d}\zeta \\ &= \left(e_s r_s e_s \right)|_0^L = \mathbf{y}^\top \mathbf{u} \end{aligned} \tag{5}$$

⋆ This work has received funding from the ANR IMPACTS project under the reference code ANR-21-CE48-0018, EIPHI Graduate School contract ANR-17-EURE-0002 and also by ANID projects FONDECYT 1191544 and BASAL FB0008..

[1] Defining the pseudo bracket

$$\{\mathcal{Z}|\mathcal{W}\} = \delta_s \mathcal{Z} \partial_\zeta \left(\delta_s \mathcal{W} \right)$$

for some smooth functionals $\mathcal{Z}$ and $\mathcal{W}$.

which shows that the BC-IPHS (3) is conservative and satisfies the first and second laws of Thermodynamics. From the isotropic assumption (*i.e.* k constant) the heat equation admits as equilibrium profiles [3] $T^* \in \mathcal{T}^*$ where

$$\mathcal{T}^* = \{T = m\zeta + b, \zeta \in [0, L], (m, b) \in \mathbb{R}\}$$

is the space of temperature equilibrium profiles satisfying $\partial_\zeta (k\partial_\zeta T^*) = 0$.

3. BOUNDARY CONTROL BY INTERCONNECTION

In this section we consider control by interconnection, *i.e.* the system is interconnected at the boundary of its spatial domain to a dynamic controller in a power preserving way as in Figure 1.

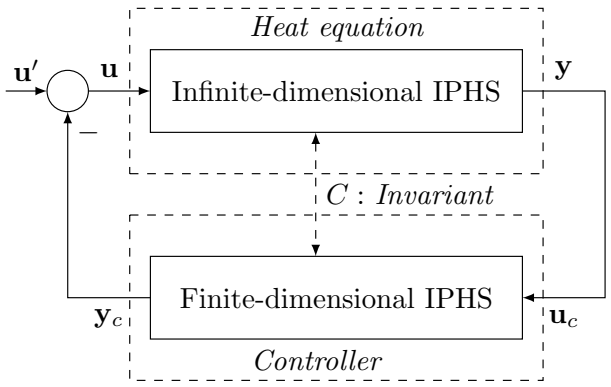


Fig. 1. CbI of the heat equation

Inspired by the classical *Control by Interconnection* method the dynamic controller is looked for under the non linear formulation

$$\dot{x}_c = G_c(x_c, \mathbf{u}_c)\mathbf{u}_c, \qquad \mathbf{y}_c = G_c^\top(x_c, \mathbf{u}_c)e_c, \tag{6}$$

where $e_c = \partial_{x_c} H_c$ and H_c a smooth function of x_c. This interconnection leads to the coupled PDE-ODE system

$$\begin{aligned} \underbrace{\begin{bmatrix} \partial_t s \\ \dot{x}_c \end{bmatrix}}_{\dot{\mathbf{x}}_{cl}} &= \underbrace{\begin{bmatrix} r_s\partial_\zeta(\cdot) + \partial_\zeta(r_s\cdot) & 0 \\ G_c(x_c, \mathbf{u}_c)W_C & 0 \end{bmatrix}}_{\mathcal{J}_{cl}} \underbrace{\begin{bmatrix} e_s \\ e_c \end{bmatrix}}_{\mathbf{e}_{cl}} \\ \mathbf{u}' &= \underbrace{\begin{bmatrix} W_B & G_c^\top(x_c, \mathbf{u}_c) \end{bmatrix}}_{W_{B_{cl}}} \mathbf{e}_{cl} \end{aligned} \tag{7}$$

where $\mathbf{e}_{cl} \in \mathcal{E}_{cl}$ denotes the vector of co-states of the closed-loop system.

Definition 1. [7, 5] Consider the boundary control system of Figure 1 with $\mathbf{u}' = 0$. A function $C : L^2([0, L], \mathbb{R}) \times \mathbb{R} \to \mathbb{R}$ is a system's invariant if $\dot{C} = 0$ along the closed loop trajectories. C is a structural invariant if $\dot{C} = 0$ along the closed loop trajectories for any $\mathbf{e}_{cl}$.

In this study we consider the following assumption.

Assumption 1. The function $C(s, x_c)$ is of the form

$$C(s, x_c) = \Gamma x_c + \int_0^L f(s(\zeta))\mathrm{d}\zeta = \kappa \tag{8}$$

where κ is a constant and $f(s) \in H^1([0, L], \mathbb{R})$ is a continuous function.

Proposition 1. Consider the BC system of Figure 1 with $\mathbf{u}' = 0$. Then (8) is a closed loop invariant if

$$\langle \mathcal{J}_{cl}\boldsymbol{\epsilon}, \mathbf{e}_{cl} \rangle = 0 \tag{9}$$

$$\begin{bmatrix} W_B & G_c^\top(x_c, \mathbf{u}_c) \end{bmatrix} \boldsymbol{\epsilon} = 0 \tag{10}$$

where

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_s \\ \epsilon_c \end{bmatrix} = \begin{bmatrix} \delta_s C \\ \partial_{x_c} C \end{bmatrix} = \begin{bmatrix} \partial_s f(s) \\ \Gamma \end{bmatrix} \tag{11}$$

In the case of the heat equation Proposition 1 leads to Proposition 2.

Proposition 2. The function C satisfies Proposition 1 if $f(s) = \alpha u(s) + c_1$ where c_1 is a function that does not depend on s. The state of the control system (6) is then given by the state feedback

$$x_c = -\frac{\alpha}{\Gamma}\int_0^L u(s)\mathrm{d}\zeta + \bar{k}/\Gamma = -\frac{\alpha}{\Gamma}\mathcal{H}(s) + \bar{k}/\Gamma \tag{12}$$

with $\bar{k} = \left(k + \int_0^L c_1 \mathrm{d}\zeta\right)$ and the controller energy function is

$$H_c = \frac{\Gamma}{\alpha}x_c + k_c = -\mathcal{H}(s) + k' \tag{13}$$

where $k' = \frac{\bar{k}}{\alpha} + k_c$, with k_c a constant. In accordance with the first principle of Thermodynamics the closed loop energy function is then constant and equal to k'.

4. ENTROPY ASSIGNMENT

In the previous section we have shown that control by interconnection with a dynamic controller of the form (6) with x_c satisfying (12) allows to preserve the closed loop energy function satisfying the first principle of Thermodynamics. In the case of the heat equation, the only physical domain that is considered is the thermal domain. In this case the internal energy is then preserved and one can use the irreversible entropy creation for control design, leading to Proposition 3.

Proposition 3. The boundary controller (6) with

$$G_c^\top = \frac{\alpha}{\Gamma}\left(\begin{bmatrix} -\frac{km^*}{T|_L} \\ -\frac{km^*}{T|_0} \end{bmatrix} + \frac{1}{c_v}\Phi(x_c)\begin{bmatrix} T(T - T^*)|_L \\ -T(T - T^*)|_0 \end{bmatrix}\right) \tag{14}$$

where $\Phi(x_c) = \Phi(x_c)^\top \geq 0$, exponentially stabilizes (3) to the desired equilibrium profile T^*.

In Proposition 3 a classical Lyapunov function of the form

$$\mathcal{V}(T, T^*) = \int_0^L \frac{1}{2}(T - T^*)^2 \mathrm{d}\zeta. \tag{15}$$

is used to design the dynamic controller (6). One can show that

$$\dot{\mathcal{V}} = -(\sigma + \mathbf{y}^\top \Xi\Phi\Xi^\top \mathbf{y}) < 0, \tag{16}$$

with $\Xi = \frac{1}{c_v}\begin{bmatrix} (T - T^*)|_L & 0 \\ 0 & (T - T^*)|_0 \end{bmatrix}$, $\Phi > 0$, $\forall T \neq T^*$ and $\dot{\mathcal{V}} = 0$ when $T = T^*$. The choice of Φ allows to shape the irreversible entropy creation in the direction of the desired equilibrium.

5. SIMULATIONS

We consider a copper bar of length $L = 0.1m$ and cross-sectional area of $10^{-4}\mathrm{m}^2$. Copper material properties can be found in [1]. We choose the following constitutive relation for the temperature $T(\zeta, t) = C_0 e^{s(\zeta,t)/c_v}$, where

C_0 is a constant such that the initial temperature profile is

$$T_0 = T(\zeta, 0) = 303.15°\text{K}, \quad \forall \zeta \in [0, 0.1]$$

Similarly, the desired temperature equilibrium profile is defined by the slope $m^* = 150°\text{K/m}$ and bias $b^* = 313.15°\text{K}$, i.e.,

$$T^* = 150\zeta + 313.15, \quad \zeta \in [0, 0.1]$$

We choose $\Phi(x_c)$ in Proposition 3 as

$$\Phi(x_c) = c_v \begin{bmatrix} \frac{a|x_c|+\phi^*}{T^2|_L} & 0 \\ 0 & \frac{a|x_c|+\phi^*}{T^2|_0} \end{bmatrix} > 0 \tag{17}$$

where $a > 0$ and $\phi^* > 0$ are constant, leading to the following control law

$$\mathbf{u} = -\frac{\Gamma}{\alpha} G_c^\top(x_c) = \begin{bmatrix} \frac{km^*}{T|_L} - \frac{T-T^*}{T}\Big|_L (a|x_c| + \phi^*) \\ \frac{km^*}{T|_0} + \frac{T-T^*}{T}\Big|_0 (a|x_c| + \phi^*) \end{bmatrix} \tag{18}$$

or equivalently by using (12) with $\alpha = -1$, $\Gamma = 1$ and $\bar{k} = -\mathcal{H}_0$

$$\mathbf{u} = \begin{bmatrix} \frac{km^*}{T|_L} - \frac{T-T^*}{T}\Big|_L \left(a \left|\int_0^L (u(\zeta,t) - u_0)\mathrm{d}\zeta\right| + \phi^*\right) \\ \frac{km^*}{T|_0} + \frac{T-T^*}{T}\Big|_0 \left(a \left|\int_0^L (u(\zeta,t) - u_0)\mathrm{d}\zeta\right| + \phi^*\right) \end{bmatrix} \tag{19}$$

The closed loop performances are shown in Figure 2 and Figure 3.

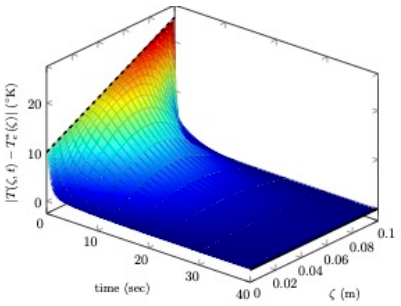


Fig. 2. Response with $\phi^* = 2$ and $a = 0.3$. Left: Absolute error $|T - T^*|$.

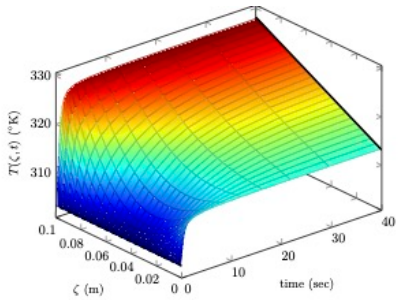


Fig. 3. Response with $\phi^* = 2$ and $a = 0.3$. Temperature trajectories.

We can see that using (19) the closed loop system is exponentially stable and reaches the desired equilibrium profile. The controller parameters can be used to modify the closed loop performances.

6. CONCLUSION

In this extended abstract we have shown how BC-IPHS formulations can be used for control design in the case of the heat equation. The main advantage of using such formulation is the physical interpretation of the closed loop system, and the design of an appropriate Lyapunov function for control design. Even if the considered case is rather simple as it corresponds to a fully actuated heat equation, it paves the way to more complex scenarios where the control is applied to only one boundary, the other being subject to a reflective condition, or unstable heat equation with distributed source terms. It should also allow to derive alternative Lyapunov functions leading to other control design with lower control efforts.

REFERENCES

[1] W.Gale, and T.Totemeier(Eds.), *Smithells Metals Reference Book*, 8th Edition, Butterworth-Heinemann, Burlington, USA, 2004.

[2] B. Jacob and H. Zwart, *Linear port-Hamiltonian Systems on Infinite- Dimensional Spaces*. Volume 223 of Operator Theory: Advances and Applications. Birkhuser/Springer Basel AG, Basel, 2012.

[3] M. Krstic and A. Smyshlyaev, *Boundary Control of PDEs: A Course on Backstepping Designs*, ser. Advances in Design and Control. 2008

[4] Y. Le Gorrec, H. Zwart and B. Maschke, *Dirac structures and Boundary Control Systems associated with Skew-Symmetric Differential Operators*. SIAM Journal on Control and Optimization, Vol. 44(5), Pages : 1864892, 2005.

[5] A. Macchelli, Y. Le Gorrec, Y., H. Ramírez, and H. Zwart, *On the synthesis of boundary control laws for distributed port-Hamiltonian systems,* Automatic Control, IEEE Transactions on, 62(4), 1700713, 2017.

[6] H. Ramirez, Y. Le Gorrec, and B. Maschke, *Boundary controlled irreversible port-Hamiltonian systems*, Chemical Engineering Science 248, 117107, 2022.

[7] A. J. van der Schaft, *L2-Gain and Passivity Techniques in Non- linear Control*, 2nd Edition, Springer-Verlag, New York, USA, 2000.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On the Computational Complexity of the Moment-SOS Hierarchy for Polynomial Optimization

Sander Gribling * **Sven Polak** ** **Lucas Slot** **

* *Université de Paris, CNRS, IRIF, F-75013, Paris, France*
** *Centrum Wiskunde & Informatica (CWI). Amsterdam, The Netherlands*

Abstract: The moment-sum-of-squares (moment-sos) hierarchy is one of the most celebrated and widely applied methods for approximating the minimum of an n-variate polynomial over a feasible region defined by polynomial (in)equalities. A key feature of the hierarchy is that it can be formulated as a semidefinite program of size polynomial in the number of variables n. Although this suggests that it may therefore be computed in polynomial time, this is not necessarily the case. Indeed, as O'Donnell [9] and later Raghavendra & Weitz [12] show, there exist examples where the sos-representations used in the hierarchy have exponential bit-complexity. We study the computational complexity of the sos-hierarchy, complementing and expanding upon earlier work of Raghavendra & Weitz [12]. In particular, we establish algebraic and geometric conditions under which polynomial-time computation is possible. As this work is still ongoing, our results should be treated as preliminary.

Keywords: polynomial optimization; sum-of-squares hierarchy; computational complexity.

1. INTRODUCTION

Consider the polynomial optimization problem:

$$\begin{aligned} f_{\min} := \min\ & f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \geq 0 \quad (1 \leq i \leq m), \\ & h_j(\mathbf{x}) = 0 \quad (1 \leq j \leq \ell), \\ & \mathbf{x} \in \mathbb{R}^n, \end{aligned} \tag{POP}$$

where $f, g_i, h_j \in \mathbb{R}[\mathbf{x}]$ are given n-variate polynomials. The feasible region of (POP) is a basic *semialgebraic* set, which we denote by:

$$S(\mathbf{g}, \mathbf{h}) := \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \geq 0,\ h_j(\mathbf{x}) = 0\}.$$

Problems of the form (POP) are generally hard and non-convex. They naturally capture several classical combinatorial problems, and have applications in finance, energy optimization, machine learning, optimal control and quantum computing. As they are often intractable, several techniques have been proposed to approximate them. Perhaps the most well-known and studied among these techniques is the so-called *moment-sum-of-squares (moment-sos) hierarchy*, due to Lasserre [5] and Parillo [10]. The main idea behind the hierarchy is that one can certify the nonnegativity of a polynomial $p \in \mathbb{R}[\mathbf{x}]$ on $S(\mathbf{g}, \mathbf{h})$ by representing it as a *weighted sum of squares*:

$$p(\mathbf{x}) = \sum_{i=0}^{m} g_i(\mathbf{x})\sigma_i(\mathbf{x}) + \sum_{j=1}^{\ell} h_j(\mathbf{x})p_j(\mathbf{x}), \tag{1}$$

where $\sigma_i \in \Sigma[\mathbf{x}]$ are sums of squares, $p_j \in \mathbb{R}[\mathbf{x}]$ and we set $g_0(\mathbf{x}) = 1$ for convenience. We say that a representation (1) is of degree t if $\deg(g_i\sigma_i) \leq t$ and $\deg(h_jp_j) \leq t$ for all i, j. For $t \in \mathbb{N}$, one then obtains a lower bound $\mathrm{sos}(f)_t \leq f_{\min}$ on the minimum of f by:

$$\mathrm{sos}(f)_t := \sup_{\lambda \in \mathbb{R}} \left\{ \lambda : \begin{array}{l} f - \lambda \text{ has a represen-} \\ \text{tation (1) of degree } 2t \end{array} \right\}. \tag{SOS}$$

Under a minor assumption on $S(\mathbf{g}, \mathbf{h})$ (see Definition 1 below), the bounds $\mathrm{sos}(f)_t$ converge to $f_{\min}$ as $t \to \infty$. The *rate* of this convergence has been the subject of some study, see, e.g., [1], [6], [7], [8], [13], [14].

For fixed level t, the lower bound $\mathrm{sos}(f)_t$ may be computed by solving a semidefinite program (SDP) involving matrices of size polynomial in n. It is often claimed that one may therefore (approximately) compute $\mathrm{sos}(f)_t$ in polynomial time, for instance by applying the ellipsoid algorithm. As was noted by O'Donnell [9] and later by Raghavendra & Weitz [12], this is not necessarily the case. Indeed, polynomial runtime of the ellipsoid algorithm is only guaranteed when the feasible region of the SDP contains an *inner ball* which is not too small, and is contained in an *outer ball* which is not too large. Informally, this means that it is possible to choose the coefficients of the multipliers σ_i, p_j in the representation (1) so that their *bit-complexity* is polynomial in n. We call such a representation *compact*.

We will consider below sets $S(\mathbf{g}, \mathbf{h})$ with the following minor algebraic boundedness condition. It is slightly stronger than the usual *Archimedean condition*, but it is more convenient to work with for our purposes.

Definition 1. We say that (POP) is *explicitly bounded* if there exists a constant $R > 0$ such that:

$$g_1(\mathbf{x}) = R - \|\mathbf{x}\|_2^2.$$

Roughly speaking, if we accept small additive errors in the solution, the condition above guarantees the existence of the inner ball (see [12]).

The remaining question, then, is whether an outer ball always exists. O'Donnell [9] shows that in fact, it does not; he constructs an example where every representation (1) of $f(\mathbf{x}) - \mathrm{sos}(f)_2$ necessarily involves multipliers σ_i, h_j whose coefficients are doubly-exponentially large in n. Raghavendra & Weitz [12] subsequently show that it is possible to construct such an example even when the equalities $\mathbf{h}$ include the boolean constraints $\mathbf{x}_i - \mathbf{x}_i^2 = 0$, negatively answering a question posed by O'Donnell. On the positive side, they show conditions under which existence of a compact representation (1) is guaranteed. These conditions are met for the reformulation of several well-known combinatorial problems as a (POP), as well as for optimization over the unit hypersphere. In order for their result to make sense, we must make the natural assumption that the coefficients of the objective function f and the polynomials g_i, h_j defining the feasible region $S(\mathbf{g}, \mathbf{h})$ of (POP) have polynomial bit-complexity.

Assumption 1. Throughout, we assume that the coefficients of the polynomials f, g_i, h_j in (POP) have polynomial bitsize in n and their degree is independent of n.

Theorem 2. (Main positive result of [12], paraphrased). Let $S(\mathbf{g}, \mathbf{h})$ be a semialgebraic set and let $t \in \mathbb{N}$ be fixed. Suppose that the following conditions are satisfied:

(1) The set $S(\mathbf{g}, \mathbf{h})$ is explicitly bounded: $g_1(\mathbf{x}) = R - \sum_{i=1}^n \mathbf{x}_i^2$ for some $R \leq 2^{\mathrm{poly}(n)}$
(2) For any $p \in \mathbb{R}[\mathbf{x}]_t$ with $p(\mathbf{x}) = 0$ for all $\mathbf{x} \in S(\mathbf{g}, \mathbf{h})$, there are $p_1, p_2, \ldots, p_\ell \in \mathbb{R}[\mathbf{x}]$ such that:
$$p(\mathbf{x}) = \sum_{j=1}^{\ell} p_j(\mathbf{x}) h_j(\mathbf{x}),$$
and $\deg(p_j h_j) = O(t)$.
(3) Let μ be the uniform probability measure on $S(\mathbf{g}, \mathbf{h})$. The *moment matrix* $M(\mu)_t$ defined by:
$$\big(M(\mu)_t\big)_{\alpha,\beta} := \int_{S(\mathbf{g},\mathbf{h})} \mathbf{x}^{\alpha+\beta} d\mu(\mathbf{x}) \quad (\alpha, \beta \in \mathbb{N}^n_t)$$
has smallest non-zero eigenvalue $\geq 2^{-\mathrm{poly}(n)}$.
(4) There exists an $\eta \geq 2^{-\mathrm{poly}(n)}$ such that $g_i(\mathbf{x}) \geq \eta$ for all $\mathbf{x} \in S(\mathbf{g}, \mathbf{h})$ and $1 \leq i \leq m$.

Then the program (SOS) has an (approximately) optimum solution involving only multipliers σ_i, p_j whose coefficients are at most $2^{\mathrm{poly}(n)}$.

1.1 Our contributions

The main goal of this paper is to carefully map under what circumstances computation of the bound $\mathrm{sos}(f)_t$ and the corresponding representation (SOS) is guaranteed to be possible in polynomial time.

Our first contribution is the following theorem.

Theorem 3. Let $S(\mathbf{g}, \mathbf{h})$ be a semialgebraic set and let $t \in \mathbb{N}$ be fixed. Suppose that the following conditions are satisfied:

(1) The set $S(\mathbf{g}, \mathbf{h})$ is explicitly bounded: $g_1(\mathbf{x}) = R - \sum_{i=1}^n \mathbf{x}_i^2$ for some $R \leq 2^{\mathrm{poly}(n)}$
(2) For any $p \in \mathbb{R}[\mathbf{x}]_{2t}$ with $p(\mathbf{x}) = 0$ for all $\mathbf{x} \in S(\mathbf{g}, \mathbf{h})$, there are $p_1, p_2, \ldots, p_\ell \in \mathbb{R}[\mathbf{x}]$ such that:
$$p(\mathbf{x}) = \sum_{j=1}^{\ell} p_j(\mathbf{x}) h_j(\mathbf{x}),$$
and $\deg(p_j h_j) \leq 2t$.
(3) There exists a constant $C \neq 0$ with $|C| \leq 2^{\mathrm{poly}(n)}$ such that the moments of the uniform probability measure μ supported on $S(\mathbf{g}, \mathbf{h})$ satisfy:
$$m(\mu)_\alpha := \int_{S(\mathbf{g},\mathbf{h})} \mathbf{x}^\alpha d\mu(\mathbf{x}) \in \tfrac{1}{C} \cdot \mathbb{N}$$
for each $\alpha \in \mathbb{N}^n_{2t}$.

Then for fixed $t \in \mathbb{N}$ and $\varepsilon \geq 2^{-\mathrm{poly}(n)}$, the bound $\mathrm{sos}(f)_t$ may be computed in polynomial time in n up to an additive error of at most ε.

Theorem 3 differs from the result of Raghavendra & Weitz in two ways: First, as we see in Section 2, it applies to several natural settings where Theorem 2 may not be applied, such as optimization over the unit ball and the standard simplex. Second, its statement is stronger in the sense that it guarantees polynomial-time computation of the bound $\mathrm{sos}(f)_t$, whereas Theorem 2 only guarantees existence of a compact representation (1).

Our second contribution is an alternative, *geometric* condition on the feasible region $S(\mathbf{g}, \mathbf{h})$ of (POP) which guarantees polynomial-time computation of $\mathrm{sos}(f)_t$ in the special case where the formulation does not contain any equality constraints.

Theorem 4. Let $S(\mathbf{g}) \subseteq \mathbb{R}^n$ be a full-dimensional, semialgebraic set defined only by inequalities. Assume that the following two conditions are satisfied:

(1) $S(\mathbf{g})$ is explicitly bounded with constant $R \leq 2^{\mathrm{poly}(n)}$, i.e., $g_1(\mathbf{x}) = R - \sum_{i=1}^n \mathbf{x}_i^2$.
(2) $S(\mathbf{g})$ contains a hypercube of size $r \geq 2^{-\mathrm{poly}(n)}$, i.e., $[-r, r]^n + z \subseteq S(\mathbf{g})$ for some $z \in \mathbb{R}^n$.

Then for fixed $t \in \mathbb{N}$ and $\varepsilon \geq 2^{-\mathrm{poly}(n)}$, the bound $\mathrm{sos}(f)_t$ may be computed in polynomial time in n up to an additive error of at most ε.

A natural motivation for the assumptions of Theorem 4 is that without them, the ellipsoid method would not be guaranteed to find a feasible point in $S(\mathbf{g})$ in polynomial time, *even if it were a convex set.* As we see below in Proposition 8, it is possible to choose the constraints g_i such that second condition of Theorem 4 is not satisfied. Interestingly, the resulting semialgebraic set $S(\mathbf{g})$ does not satisfy the conditions of Theorem 2 or Theorem 3, either.

Finally, we make explicit the connection between computational aspects of the primal formulation (SOS) of the sos-hierarchy, and its dual formulation (MOM) in terms of *moments* (see below). This connection is implicitly present in the proof of Theorem 2 in [12].

Theorem 5. Let $S(\mathbf{g}, \mathbf{h})$ be a semialgebraic set and suppose that the conditions of Theorem 3 or Theorem 4 are satisfied. Then one may (approximately) compute the bound $\mathrm{sos}(f)_t$ in time polynomial in n, using either its primal formulation (SOS) or its dual formulation (MOM). In particular, there then exists a sum-of-squares representation (1) proving nonnegativity of $f - \mathrm{sos}(f)_t + \varepsilon$ on $S(\mathbf{g}, \mathbf{h})$ with low bit-complexity.

2. SOME REMARKS AND APPLICATIONS

Here, we give some small examples that illustrate the advantages and limitations of our results with respect to previous work.

Example 6. The unit ball $B^n \subseteq \mathbb{R}^n$ and the standard simplex $\Delta^n \subseteq \mathbb{R}^n$ are semialgebraic sets, defined by:

$$B^n = \{\mathbf{x} \in \mathbb{R}^n : 1 - \|\mathbf{x}\|_2^2 \geq 0\},$$
$$\Delta^n = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_i \geq 0,\ 1 - \textstyle\sum_{i=1}^n \mathbf{x}_i \geq 0\}.$$

It is straightforward to see that they both satisfy the conditions of Theorem 4. They do not, however, satisfy the third condition of Theorem 2.

Remark 7. In general, our Theorem 3 and Theorem 4 are better equipped to deal with non-discrete semialgebraic sets $S(\mathbf{g}, \mathbf{h})$ than Theorem 2. The fourth condition of Theorem 2, which demands in particular that $g_i(\mathbf{x}) > 0$ for each $\mathbf{x} \in S(\mathbf{g}, \mathbf{h})$, is generally quite hard to satisfy. For instance, the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n : \|\mathbf{x}\|_2^2 = 1\}$ satisfies this condition (as it is defined without using any inequalities), but it no longer does so after adding a linear constraint such as $\mathbf{x}_1 \geq 0$. On the other hand, our Theorem 3 still applies.

The next proposition shows that the second condition of Theorem 4 is not superfluous, via a simple repeated-squaring argument.

Proposition 8. There exists a full-dimensional semialgebraic set $S(\mathbf{g})$, defined only by polynomial inequalities $\mathbf{g} = (g_1, g_2, \ldots, g_m)$ whose coefficients have bit complexity polynomial in n, which does not contain a (translated) cube $[-r, r]^n$ for $r \geq 2^{-\mathrm{poly}(n)}$.

Proof. Let $S(\mathbf{g})$ be the set defined by the system of inequalities:

$$\begin{aligned} \mathbf{x}_i &\geq 0 \qquad (1 \leq i \leq n), \\ \mathbf{x}_i - \mathbf{x}_{i+1}^2 &\leq 0 \qquad (1 \leq i \leq n-1), \\ \mathbf{x}_n &\leq 1/2. \end{aligned}$$

Set $r := 2^{-2^n}$. Then $[0, r]^n \subseteq S(\mathbf{g})$ and so $S(\mathbf{g})$ is full-dimensional. But from the inequalities it follows that $0 \leq \mathbf{x}_1 \leq r$ for any $x \in S(\mathbf{g})$, meaning $S(\mathbf{g})$ cannot contain a (translated) cube of size $2^{-\mathrm{poly}(n)}$.

It is instructive to note that the set $S(\mathbf{g})$ of Proposition 8 does not satisfy the conditions of Theorem 2 and Theorem 3. Indeed, we may easily compute the moments for the uniform probability measure μ on $S(\mathbf{g})$ and observe that:

$$\int_{S(\mathbf{g})} \mathbf{x}_1^2 d\mu(\mathbf{x}) = \int_{S(\mathbf{g})} \mathbf{x}_1^2 d\mathbf{x} \Big/ \int_{S(\mathbf{g})} 1 d\mathbf{x} \leq 2^{-2^n}$$

In particular, the smallest nonzero eigenvalue of the moment matrix $M(\mu)_t$ is at most 2^{-2^n} for any $t \geq 1$.

3. OUTLINE OF THE PROOFS

It will be convenient to work with the dual formulation of (SOS), which reads (see, e.g., [2]):

$$\begin{aligned} \mathrm{mom}(f)_t := \inf\ & L(f) \\ \text{s.t. } & L(1) = 1, \\ & L(g_i p^2) \geq 0, \quad (g_i p^2 \in \mathbb{R}[\mathbf{x}]_{2t}) \\ & L(h_j \mathbf{x}^\alpha) = 0, \ \ (h_j \mathbf{x}^\alpha \in \mathbb{R}[\mathbf{x}]_{2t}) \\ & L \in \mathbb{R}[\mathbf{x}]_{2t}^*. \end{aligned} \tag{MOM}$$

Under the assumption of explicit boundedness, these formulations are actually equivalent.

Theorem 9. ([4]). If (POP) is explicitly bounded, we have strong duality between the primal and dual formulations (SOS) and (MOM) of the moment-sos hierarchy. That is, we then have:

$$\mathrm{sos}(f)_t = \mathrm{mom}(f)_t \quad \forall t \in \mathbb{N}.$$

As we alluded to in the introduction, explicit boundedness of $S(\mathbf{g}, \mathbf{h})$ also gives a bound on the feasible region of (MOM).

Lemma 10. (see, e.g., [11]). Assume that $S(\mathbf{g}, \mathbf{h})$ is explicitly bounded for some $R > 0$. Let $L \in \mathbb{R}[x]_{2t}^*$ be a feasible solution to (MOM). Then $|L(\mathbf{x}^\alpha)| \leq R^{|\alpha|}$ for all $\alpha \in \mathbb{N}_{2t}^n$.

There is a natural relation between the dual formulation (MOM) and *moments* of measures supported on $S(\mathbf{g}, \mathbf{h})$, which clarifies the assumptions made in Theorem 2 and Theorem 3. For a measure μ supported on $S(\mathbf{g}, \mathbf{h})$, the *moment* of degree $\alpha \in \mathbb{N}^n$ is defined by:

$$m(\mu)_\alpha := \int_{S(\mathbf{g},\mathbf{h})} \mathbf{x}^\alpha d\mu(\mathbf{x}).$$

For $t \in \mathbb{N}$, the (truncated) moment matrix $M(\mu)_t$ of order t for μ is then given by:

$$\big(M(\mu)_t\big)_{\alpha,\beta} = m(\mu)_{\alpha+\beta}. \tag{2}$$

Consider the linear functional $L_\mu \in \mathbb{R}[\mathbf{x}]_{2t}^*$ defined by:

$$L_\mu(p) := \int_{S(\mathbf{g},\mathbf{h})} p(\mathbf{x}) d\mu(x) \quad (p \in \mathbb{R}[\mathbf{x}]_{2t}).$$

For any constraint g_i and $p \in \mathbb{R}[\mathbf{x}]$ with $\deg(g_i p^2) \leq 2t$, we have:

$$L_\mu(g_i p^2) = \mathbf{p}^\top M(g_i \mu)_t \mathbf{p} = \int_{S(\mathbf{g},\mathbf{h})} p^2(\mathbf{x}) g_i(\mathbf{x}) d\mu(\mathbf{x}) \geq 0,$$

where $\mathbf{p}$ denotes the vector of coefficients of $p \in \mathbb{R}[\mathbf{x}]_t$ in the monomial basis. Here the (α, β)-entry of the *localizing matrix* $M(g_i\mu)_t$ is defined as $\int_{S(\mathbf{g},\mathbf{h})} g_i(\mathbf{x}) \mathbf{x}^{\alpha+\beta} d\mu(\mathbf{x})$. In particular, for each i the matrix $M(g_i\mu)_t$ is positive semidefinite. Furthermore, for any constraint h_j and $\alpha \in \mathbb{N}^n$ with $\deg(\mathbf{x}^\alpha h_j) \leq 2t$, we have:

$$L_\mu(h_j \mathbf{x}^\alpha) = \int_{S(\mathbf{g},\mathbf{h})} h_j(\mathbf{x}) \mathbf{x}^\alpha d\mu(\mathbf{x}) = 0.$$

If μ is a probability measure, we get $L_\mu(1) = 1$, and it follows that L_μ is a feasible solution to (MOM).

Remark 11. The upshot is that in order to show that the feasible region of (MOM) contains an inner ball, it is enough to exhibit a probability measure μ on $S(\mathbf{g}, \mathbf{h})$ with:

(1) The smallest non-zero eigenvalue of $M(g_i\mu)_t$ is at least $2^{-\mathrm{poly}(n)}$ for each $i = 0, 1, \ldots, m$;
(2) For any $p \in \mathbb{R}[\mathbf{x}]_t$ with $M(\mu)_t \mathbf{p} = 0$, we have that $p(\mathbf{x}) = \sum_{j=1}^\ell h_j(\mathbf{x}) p_j(\mathbf{x})$, where $\deg(h_j p_j) \leq 2t$.

One may then use a fairly standard result on SDP complexity to conclude polynomial-time (approximate) computability of $\mathrm{mom}(f)_t$, see for instance Theorem 1.1 in [3].

3.1 Proof sketch for Theorem 3

Our proof uses similar ideas to that of Theorem 2 in [12]. Let μ be the uniform probability measure on $S(\mathbf{g}, \mathbf{h})$. We show μ satisfies the properties (1) and (2) in Remark 11. For (1), note that the third condition of Theorem 3 tells us that $C \cdot M(\mu)_t$ is an integer matrix for some $C \neq 0$, $|C| \leq 2^{\text{poly}(n)}$. The eigenvalues of integer matrices can be controlled in terms of their largest entry. By Lemma 10, we have an upper bound on the entries of $C \cdot M(\mu)_t$, and this allows us to show that the smallest nonzero eigenvalue of $M(\mu)_t$ is at least $2^{-\text{poly}(n)}$. The entries of the matrices $M(g_i\mu)_t$ are *linear combinations* of the entries of $M(\mu)_t$, involving the coefficients of the constraints g_i. As these coefficients have polynomial bit-complexity, a similar argument allows us to lower bound the smallest nonzero eigenvalue of each $M(g_i\mu)_t$. For (2), note that this is implied immediately by the second condition in Theorem 3.

3.2 Proof sketch for Theorem 4

Let μ be the normalized Lebesgue measure on $S(\mathbf{g})$. If $S(\mathbf{g})$ contains a (translated) hypercube $[-r, r]^n + z$ of size $r \geq 2^{\text{poly}(n)}$, μ trivially satisfies condition (2) in Remark 11 (as $M(\mu)_t\mathbf{p} = 0 \implies \mathbf{p} = 0$). To see that it also satisfies condition (1), we proceed as follows:

- Show that we may assume w.l.o.g. that $z = 0$, using the fact that $\|z\|_2^2 \leq R \leq 2^{\text{poly}(n)}$ as $S(\mathbf{g})$ is explicitly bounded.
- Use known formulas (c.f. [2]) for the moments of the hypercube to show that there exists a $C \neq 0$, $|C| \leq 2^{\text{poly}(n)}$ such that $C \cdot M(\mu')_t$ is integer, where μ' is the restriction of μ to $[-r, r]^n$.
- Conclude that the smallest eigenvalue of $M(g_i\mu')_t$ is at least $2^{-\text{poly}(n)}$ for each i.
- Finally, use the fact that $[-r, r]^n \subseteq S(\mathbf{g})$ to conclude that $M(g_i\mu)_t \succeq M(g_i\mu')_t$ for all i, finishing the proof.

3.3 Proof sketch for Theorem 5

In our proofs of Theorem 3 and Theorem 4, we show that the dual formulation (MOM) of the sos-hierarchy may be (approximately) solved in polynomial time using (e.g.) the ellipsoid algorithm. This does not immediately imply that the same is true for the primal formulation (SOS). By considering the pairing between primal feasible solutions (i.e., representations of the form (1)) and the dual feasible solution L_μ used in our proofs, we show that the feasible region of the primal formulation is also bounded. This pairing also plays an important role in the proof of Theorem 2 in [12]. Indeed, for any representation $f(\mathbf{x}) = \sum_i g_i(\mathbf{x})\sigma_i(\mathbf{x}) + \sum_j h_j(\mathbf{x})p_j(\mathbf{x})$, we have:

$$L_\mu(\sum_i g_i\sigma_i + \sum_j h_j p_j) = L_\mu(f) \leq 2^{\text{poly}(n)}.$$

This allows us to bound $L_\mu(g_i\sigma_i)$, which in turn allows us to bound the size of the coefficients of σ_i. To bound the coefficients of the p_j one uses the second condition of Theorem 3.

REFERENCES

[1] Lorenzo Baldi and Bernard Mourrain. On moment approximation and the effective Putinar's Positivstellensatz, 2021.

[2] Etienne de Klerk and Monique Laurent. *A Survey of Semidefinite Programming Approaches to the Generalized Problem of Moments and Their Error Analysis*, pages 17–56. Springer International Publishing, Cham, 2019.

[3] Etienne de Klerk and Frank Vallentin. On the turing model complexity of interior point methods for semidefinite programming. *SIAM Journal on Optimization*, 26(3):1944–1961, 2016.

[4] Cédric Josz and Didier Henrion. Strong duality in Lasserre's hierarchy for polynomial optimization. *Opt. Lett.*, 10:3–10, 2016.

[5] Jean-Bernard Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[6] Monique Laurent and Lucas Slot. An effective version of Schmüdgen's Positvstellensatz for the hypercube. *arXiv: 2109. 09528*, 2021.

[7] Ngoc Hoang Anh Mai and Victor Magron. On the complexity of Putinar-Vasilescu's Positivstellensatz. *arXiv: 2104. 11606*, 2021.

[8] Jiawang Nie and Markus Schweighofer. On the complexity of Putinar's Positivstellensatz. *Journal of Complexity*, 23:135–150, 2007.

[9] Ryan O'Donnell. SOS Is Not Obviously Automatizable, Even Approximately. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 59:1–59:10, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[10] Pablo Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming Series B*, 96:293–320, 2003.

[11] Stefano Pironio, Miguel Navascués, and Antonio Acín. Convergent relaxations of polynomial optimization problems with noncommuting variables. *SIAM J. Optim.*, 20:2157–2180, 2010.

[12] Prasad Raghavendra and Benjamin Weitz. On the Bit Complexity of Sum-of-Squares Proofs. *ICALP*, 80:1–13, 2017.

[13] Markus Schweighofer. On the complexity of Schmüdgen's Positivstellensatz. *Journal of Complexity*, 20(4):529–543, 2004.

[14] Lucas Slot and Monique Laurent. Sum-of-squares hierarchies for binary polynomial optimization. In Mohit Singh and David P. Williamson, editors, *Integer Programming and Combinatorial Optimization*, pages 43–57, Cham, 2021. Springer International Publishing.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Future Control Community Papers on Algorithmic Trading Should Pay More Attention to Backtesting

B. Ross Barmish

Robust Trading Solutions, LLC
Boxford, MA 01921
email: bob.barmish@gmail.com

Abstract: The takeoff point for this work is the emerging body of literature which addresses algorithmic trading in the framework of feedback control systems. In this setting, the buying and selling of equities period is governed by the action of a controller, using past history, to determine the time-varying investment level. Almost all of the papers to date begin with a underlying mathematical model structure for the stock-price dynamics and "theoretical" performance is studied. In many cases, the parameters of the price model are not assumed to be known in advance; they are estimated over time from the realized price path. In the literature, we also see many variations on this theme. For example, the investment-level controller may have no explicit reliance on an assumed price model and instead are updated by performance variable data such as account market value or gains and losses over time. Subsequently, the authors of papers along these lines demonstrate the performance of their trading algorithms using various criteria. Given this context, we draw attention to use of the word "should" in the title. This is intentional because this short paper is an opinion piece; it does not contain new results. Instead, arguments are given that the control community will be well served if future papers devote greater attention to backtesting and standardization of benchmark data sets. It is argued that this will enable the results of one researcher to more easily replicated and compared against those of another and, in turn, this will increase the impact of control-theoretic papers on researchers and practitioners outside the control field. While it is true that a number of the control-inspired papers to date already include some backtest results, the use of widely varying data sets makes evaluation of worthiness of their "controller recipes" difficult or impossible.

Keywords: financial engineering, algorithmic stock trading, stochastic systems, backtesting

1. INTRODUCTION

The basic idea of viewing stock trading in a control-theoretic setting goes back about fifty years to the finance community where optimal control concepts were brought into play in a portfolio optimization context; e.g., see [1] and [2]. Over the last two decades, we see a number of papers bringing many other aspects of modern control theory to the fore. Among the earliest of these, [3] and [4], appear in 2001 and 2002. In the ensuing twenty years, a large number of papers along these lines have followed. To keep the length of this exposition within page-limit requirement, we refer the reader to papers [5] and [6] and their bibliographies where this literature is broadly surveyed covering the period up until 2016. Beyond 2016, we see a number of new ideas being introduced into this line of research. To mention just a few examples, in [7], [8] and [9], various generalizations of the so-called SLS method are studied, in [10], the use of modulated controllers aimed at drawdown reduction are introduced, in [11], approaches to trading in a model-predictive control context are studied and in [12], trend following from an H^∞ tracking point of view is investigated.

1.1 The Basic Setup for Control-Inspired Stock Trading

The feedback control approach to stock trading is easily understood in terms of the closed-loop feedback configuration given in Figure 1. Although this setup is not explicitly given in all of the cited literature above, it is central to almost all control-inspired research on stock trading which one can imagine.

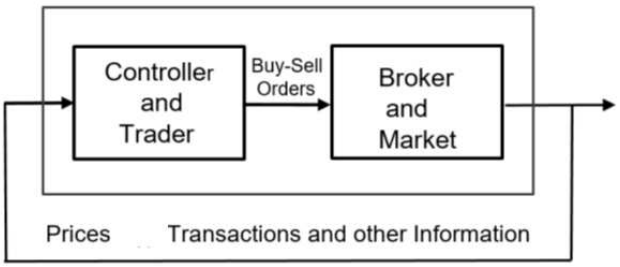


Figure 1: Feedback Loop Involving Trader and Broker

The paradigm associated with the block diagram in the figure is central to the study of a wide variety of feedback control laws which can be used to modify the time-varying investment level. In some papers, a model for future stock prices plays an important role and in other cases no stock price model is used at all and trading signals are generated based on other considerations such as the market value of the account or gains and losses over time.

1.2 This Author's Opinion of Existing Research

When looking at the existing results over the last twenty years coming from the control community, in many papers, including those of this author, there is inadequate attention being paid to "meaningful" backtesting. Too much of center stage is occupied by theoretical issues and backtest simulations which are given are of often of questionable worth. In many cases when such simulations are included, the data set is often "nonstandard" or covers too short a time duration which does not include both bull and bear markets. Simply put, it is felt that many of the backtest simulations conducted by the control field researchers are insufficient to determine whether the trading scheme being espoused will be efficacious. To illustrate the case being made in this paper, we ask the following question: Based on highly successful backtests for Tesla and Apple stock over the year 2021, can any meaningful conclusions be made about the excellence of the trading algorithm being used? Of course not. In the finance literature, the issue of "data snooping" often arises in connection with a belief that a given investment-level control algorithm is being made to look "better than deserved." In the control field, care is required to avoid cherry picking data (choice of equities, time period, etc.) and "twiddling" with parameters such as feedback gains entering into the simulation. As suggested in the sequel, the control field should adopt *standard benchmark data sets* and rules for algorithm parameter initialization and adjustment so that backtest results reported in the literature are **repeatable**and new algorithms are more easily evaluated.

2. STANDARDIZATION OF BACKTESTING

The view of this author is that the control field would be well served if papers included a credible backtest which is not easily subject to challenge. In this regard, prior to dealing with a diversified portfolio of stocks, a simulation should be carried out using some well-known benchmark such as the Dow Jones Industrial Average (DJIA) or the Standard and Poor's (S&P) 500. In addition, a sufficiently long multi-year time window should be used which includes both bull and bear markets. Finally, the performance of the algorithm over sub-windows should be also be considered. This is important because "ride quality" is very important to many traders. It does not suffice to claim that an algorithm leads to a great end result if the market value of the account undergoes large roller coaster deviations along the way. One excellent example illustrating high-quality backtesting is reference [13]. Trades are triggered based on the stock price crossing a prescribed multi-day moving average. In lieu of making the case for the "excellence" of their results based on mathematical theory on how moving averages behave, the authors provide a backtest using historical data for the DJIA covering nearly ninety years. This shifts the debate away from the quality of the mathematical justification to real-world implications such as the longstanding theory that markets are efficient and that "abnormal" rates of return, when risk-adjusted, cannot be consistently obtained.

3. FUTURE CONTROL COMMUNITY PAPERS

It is suggested that in our future papers, after the introduction and a literature review, the author provides the investment-level controller recipe and a simple backtest with as little supporting mathematical detail as possible. While it is true that some mathematical formulae will be needed to describe the controller, the theorems, proofs and price modelling can be relegated to a later section of the paper such as an appendix. Accompanying this recipe, perhaps some authors can include emphasis on the intuitive appeal of their algorithms. Then, leave it to fellow researchers to decide whether the backtest is "sufficiently impressive" to merit careful study the supporting theory. This author's opinion, controversial as it may be, is summarized as follows: *It does not suffice to provide excellent theory with weak supporting backtests. However, it does suffice to provide excellent backtests with weak supporting theory.*

Per discussion above, as in reference [13] where the Dow Jones Industrial Average is used, the view of this author, as previously mentioned, is that the chosen data set should be standard and well known to almost any possible trader. To this end, as discussed below, one very good possibility for use would be the time series for daily S&P 500 prices over a long time period. The associated Exchange Traded Fund (ETF), having ticker SPY, can be readily downloaded from Yahoo Finance. Given the high liquidity of this ETF, it is reasonable to assume in simulations that trading can be carried out with minimal friction. We also emphasize the words "long time period" above. Many types of behavior such as bull markets, bear markets and corrections should be covered.

3.1 Other Backtest Considerations

Many papers in the control literature pay little or no attention to important issues such as margin requirements and associated leverage allowed by the broker, slippage associated with the bid-ask spread and attainability of specified orders such as "market on close." When conducting backtest simulations, since one of our main objectives is to determine if promising control algorithms perform well when subjected to these practical market considerations, attention to these issues in simulation is important.

4. ON EVALUATION OF BACKTEST RESULTS

After a backtest simulation is performed, the following question arises: How do we evaluate the quality of the results? For example, if one has a 200% one-year return on some mining stock but a loss of 50% after the first six months, is this considered a good result? Some traders would be perfectly comfortable with this type of roller coaster ride but others, particularly those who are risk-averse, would not. To avoid controversy whether this result is good or bad, the suggestion in this paper is to use the method of many mutual funds to display results. That is, one begins a simulation with a fictitious $10,000 and simply provides a plot of the evolution of market value over time. For this hypothetical scenario, the 50% dip and 200% rise above will be immediately visible in the plot and each individual can judge the performance based on his or her individual utility function. While it is perfectly acceptable to report on various features of the plot such as annual returns, and highs and lows, no claims of "optimality" are appropriate. Each "examiner" of the market value plot has all the information required to make an individual judgment; see the example below for additional details.

5. EXAMPLE: TWENTY YEARS OF THE S&P 500

To make the ideas above more concrete, we imagine a control community researcher who develops a "secret" trading algorithm and conducts a 20-year backtest using price data for SPY, the ETF tracking the S&P 500. Trading begins mid-February of 1998 with the market value of the account initialized at $10,000. On day k, the trading algorithm processes the past history and determines investment level $I(k)$. Assuming short selling and use of leverage is disallowed, using $V(k)$ to denote the associated 5000+ market values of the account, the condition $0 \leq I(k) \leq V(k)$ must be satisfied. In addition, we assume that all transactions occur at the daily closing price and, for pedagogical simplicity, we neglect the small dividends which are "cast off" by the stocks comprising the index.

5.1 Evaluation: A Picture is Worth a Thousand Words

For the scenario above, in Figure 1, the black market-value plot represents the benchmark. A buy-and-hold trader sees wildly gyrating returns but ends up with about a 35% twenty-year return. On the other hand, the red plot, representing the returns on the trader's algorithm only ends up with about a 17.5% gain for the same period. Given these two plots, there is no need for the control researcher to provide an "opinion" which market-value plot is preferred; the two plots speaks for themselves. A risk-averse investor who finds the large intermediate drawdowns on the S&P 500 intolerable may give the overall market-value plot a low performance rating. The slow steady rise of the red market value associated with the algorithm only leads to a return of about 17.5% but the ride quality may result in a much higher performance rating by many more conservative traders who are unwilling to ride out a storm.

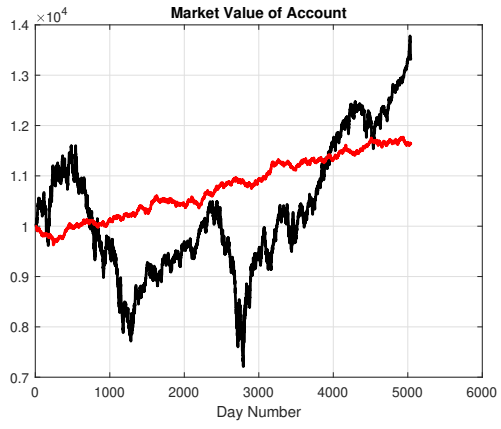


Figure 1: Growth of $10,000: Strategy Versus Benchmark

6. CONCLUSION

In this paper, it was argued that future research on algorithmic trading in the control community should pay much more attention to backtesting and use a benchmark which is simple, widely understood and covers a suitably long time period. It is the opinion of this author that the market-value plot is typically much more informative and insightful than the lengthy statistical analyses found in many papers coming from the finance community. A final point to mention is the following view held by the author: A trading algorithm which cannot be theoretically justified in a theorem-proof context, should not be disqualified from being seriously studied in a backtesting context. A backtest which is sufficiently compelling can be very important even if the trading algorithm recipe appears to be ad hoc. This includes controller recipes which may involve heuristics or use of ideas from areas such as machine learning or neural networks; e.g., see [14].

REFERENCES

[1] R. C. Merton, Lifetime Portfolio Selection Under Uncertainty: The Continuous Time Case, *Review of Economics and Statistics*,vol. 51, pp. 247-257, 1969.

[2] P. A. Samuelson, Lifetime Portfolio Selection By Dynamic Stochastic Programming, *Review of Economics and Statistics*,vol. 51, pp. 239-246, 1969.

[3] Q. Zhang, "Stock Trading: An Optimal Selling Rule," *SIAM Journal of Control Optimization*, vol. 40, pp. 64–87, 2001.

[4] N. G. Dokuchaev and A. V. Savkin, "A Bounded Risk Strategy for a Market with Non-Observable Parameters," Insurance: Mathematical Economics, vol. 30, no. 2, pp. 243–254, 2002.

[5] B. R. Barmish, J. A. Primbs, S. and S. Warnick, "On the Basics for Simulation of Feedback-Based Stock Trading Strategies," *Proceedings of the IEEE Conference on Decision and Control*, pp. 7181-7186, Florence, Italy, 2013.

[6] B. R. Barmish and J. A. Primbs, "On a New Paradigm for Stock Trading Via a Model-Free Feedback Controller," *IEEE Transactions on Automatic Control*, AC-61, pp. 662-676, 2016.

[7] M. H. Baumann, "On Stock Trading Via Feedback Control When Underlying Stock Returns Are Discontinuous," *IEEE Transactions on Automatic Control*, AC-62, pp. 2987–2992, 2017.

[8] S. Malekpour, J. A. Primbs and B. R. Barmish, "A Generalization of Simultaneous Long–Short Stock Trading to PI Controllers," *IEEE, Transactions on Automatic Control*, AC-63, pp. 3531-3536, 2018

[9] J. D. O'Brien, M. Burke and K. Burke, "A Generalized Framework for Simultaneous Long-Short Feedback Trading," *IEEE Transaction on Automatic Control*, AC-66, pp. 2652-2653, 2021.

[10] C. H. Hsieh and B. R. Barmish, "On Drawdown-Modulated Feedback in Stock Trading," *Proceedings of the IFAC World Congress*, pp. 952-958, Toulouse, France, 2017.

[11] V. Dombrovskii, T. Obyedko and M. Samorodova, "Model Predictive Control of Constrained Markovian Jump Nonlinear Stochastic Systems and Portfolio Optimization Under Market Frictions" *Automatica*, vol. 87, pp. 61-68, 2018.

[12] G. Maroni, S. Formentin, and F. Previdi, "A Robust Design Strategy for Stock Trading via Feedback Control," *Proceedings of the European Control Conference*, pp. 447–452, Naples, Italy, 2019.

[13] W. Brock, J. Lakonishok, and B. LeBaron, "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns," *The Journal of Finance*, vol. 47, pp. 1731-1764, 1992.

[14] P. D. McNelis, *Neural Networks in Finance: Gaining Predictive Edge in the Market*, Academic Press, 2005.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Loop Shaping with Scaled Relative Graphs

Thomas Chaffey * **Fulvio Forni** * **Rodolphe Sepulchre** *

* *University of Cambridge, Department of Engineering, Trumpington Street, Cambridge CB2 1PZ,* {tlc37, ff286, rs771}@cam.ac.uk.

Abstract: The Scaled Relative Graph (SRG) is a generalization of the Nyquist diagram that may be plotted for nonlinear operators, and allows nonlinear robustness margins to be defined graphically. This abstract explores techniques for shaping the SRG of an operator in order to maximize these robustness margins.

Keywords: Scaled Relative Graph, Nyquist, loop shaping, robustness

1. INTRODUCTION

Loop shaping is one of the earliest methods of controller design, originating in the work of Nyquist, Bode, Nichols and Horowitz on feedback amplifiers (Bode, 1960). The basic principle of loop shaping is to tune a system's closed loop performance by adjusting the open loop frequency response. Robustness is captured by the distance of the Nyquist diagram from the point -1; closed loop performance is captured by the sensitivity and related transfer functions. Loop shaping is still widely used in industry today – the graphical nature of the tool gives a clear view of the design tradeoffs between performance and robustness. Even in the age of modern optimal and robust control, loop shaping remains a core tool for the control engineer. The idea of enlarging stability margins eventually led to the Zames' formulation of H_∞ control (Zames, 1981), and some of the most successful methods of robust control combine H_∞ control with classical loop shaping ideas (Vinnicombe, 2000; McFarlane and Glover, 1992).

The Scaled Relative Graph (SRG) is a graphical representation of a nonlinear operator, recently introduced in the theory of optimization by Ryu et al. (2021). The SRG allows simple, intuitive proofs of convergence for optimization algorithms, and allows optimal convergence rates to be visualized as distances on a plot. The authors have recently connected the SRG to classical control theory, showing that it generalizes the Nyquist diagram of an LTI transfer function (Chaffey et al., 2021). A range of incremental stability results, including the Nyquist and circle criteria, small gain and passivity theorems and secant condition, can be interpreted as guaranteeing the separation of the SRGs of two systems in feedback, and this interpretation has led to new conditions for incremental stability (Chaffey, 2022). The distance between the two SRGs is an incremental disc margin, the reciprocal of which bounds the incremental gain of the closed loop. The SRG makes the design intuition afforded by the Nyquist diagram available for nonlinear systems.

This abstract describes ongoing research into the use of SRGs for loop shaping nonlinear feedback systems. It has long been observed that introducing nonlinearity can overcome fundamental limitations of LTI control – for example, the describing function of the Clegg integrator has a phase lag of only 38°, rather than the usual 90° of a linear integrator (Clegg, 1958). This motivates a better understanding of how nonlinearities may be used to shape the performance of a feedback system.

2. REVIEW OF SCALED RELATIVE GRAPHS

We begin this extended abstract with a brief review of the theory of SRGs.

2.1 Signal Spaces

We describe systems using operators, possibly multi-valued, on a Hilbert space. A Hilbert space $\mathcal{H}$ is a vector space equipped with an inner product, $\langle \cdot | \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$, and the induced norm $\|x\| \coloneqq \sqrt{\langle x | x \rangle}$.

We will pay particular attention to the Lebesgue space L_2. Given $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, $L_2^n(\mathbb{F})$ is defined as the set of signals $u : \mathbb{R}_{\geq 0} \to \mathbb{F}^n$ such that

$$\|u\| \coloneqq \left(\int_0^\infty u(t) \bar{u}(t) \, \mathrm{d}t \right)^{\frac{1}{2}} < \infty,$$

where $\bar{u}(t)$ denotes the conjugate transpose of $u(t)$. The inner product of $u, y \in L_2^n(\mathbb{F})$ is defined by

$$\langle u | y \rangle \coloneqq \int_0^\infty u(t) \bar{y}(t) \, \mathrm{d}t \, .$$

The Fourier transform of $u \in L_2^n(\mathbb{F})$ is defined as

$$\hat{u}(j\omega) \coloneqq \int_0^\infty e^{-j\omega t} u(t) \, \mathrm{d}t \, .$$

We omit the dimension and field when they are immaterial or clear from context.

2.2 Relations

An *operator*, or *system*, on a space $\mathcal{H}$, is a possibly multi-valued map $R : \mathcal{H} \to \mathcal{H}$. The identity operator, which maps

* The research leading to these results has received funding from the European Research Council under the Advanced ERC Grant Agreement Switchlet n. 670645, and from the Cambridge Philosophical Society.

$u \in \mathcal{H}$ to itself, is denoted by I. The *graph*, or *relation*, of an operator, is the set $\{u, y \mid u \in \operatorname{dom} R, y \in R(u)\} \subseteq \mathcal{H} \times \mathcal{H}$. We use the notions of an operator and its relation interchangeably, and denote them in the same way.

The usual operations on functions can be extended to relations. Let R and S be relations on an arbitrary Hilbert space. Then:

$$\begin{aligned} S^{-1} &= \{(y,u) \mid y \in S(u)\} \\ S + R &= \{(x, y+z) \mid (x,y) \in S, (x,z) \in R\} \\ SR &= \{(x,z) \mid \exists\, y \text{ s.t. } (x,y) \in R, (y,z) \in S\}. \end{aligned}$$

Note that S^{-1} always exists, but is not an inverse in the usual sense. In particular, it is in general not the case that $S^{-1}S = I$. The relational inverse plays a fundamental role in the techniques described in this abstract. Rather than directly shape the performance of a negative feedback interconnection, we will shape the performance of its inverse relation – a parallel interconnection.

These operations will also be used on sets of operators, with the meaning that the operations are applied elementwise to the sets (under the implicit assumption that the operators have compatible domains and codomains).

2.3 Scaled Relative Graphs

We define SRGs in the same way as Ryu et al. (2021), with the minor modification of allowing complex valued inner products.

Let $\mathcal{H}$ be a Hilbert space. The angle between $u, y \in \mathcal{H}$ is defined as

$$\angle(u,y) \coloneqq \operatorname{acos} \frac{\operatorname{Re}\langle u|y\rangle}{\|u\|\|y\|}.$$

Let $R : \mathcal{H} \to \mathcal{H}$ be an operator. Given $u_1, u_2 \in \mathcal{U} \subseteq \mathcal{H}$, $u_1 \neq u_2$, define the set of complex numbers $z_R(u_1,u_2)$ by

$$z_R(u_1,u_2) \coloneqq \left\{ \frac{\|y_1 - y_2\|}{\|u_1 - u_2\|} e^{\pm j \angle(u_1-u_2, y_1-y_2)} \;\middle|\; y_1 \in R(u_1), y_2 \in R(u_2) \right\}.$$

If $u_1 = u_2$ and there are corresponding outputs $y_1 \neq y_2$, then $z_R(u_1,u_2)$ is defined to be $\{\infty\}$. If R is single valued at u_1, $z_R(u_1,u_1)$ is the empty set.

The *Scaled Relative Graph* (SRG) of R over $\mathcal{U} \subseteq \mathcal{H}$ is then given by

$$\operatorname{SRG}_{\mathcal{U}}(R) \coloneqq \bigcup_{u_1,u_2 \in \mathcal{U}} z_R(u_1,u_2).$$

If $\mathcal{U} = \mathcal{H}$, we write $\operatorname{SRG}(R) \coloneqq \operatorname{SRG}_{\mathcal{H}}(R)$. The SRG of a class of operators is defined to be the union of their individual SRGs. Some examples of SRGs are shown in Figure 1, (a), (b) and (c).

2.4 Interconnections

The power of SRGs lies in the elegant interconnection theory of Ryu et al. (2021). Given the SRGs of two systems, the SRG of their interconnection can be bounded using simple graphical rules. Given two systems R and S, and subject to mild conditions, we have:

$$\begin{aligned} \operatorname{SRG}(\alpha R) = \operatorname{SRG}(R\alpha) &= \alpha \operatorname{SRG}(R) \\ \operatorname{SRG}(R+S) &\subseteq \operatorname{SRG}(R) + \operatorname{SRG}(S) \\ \operatorname{SRG}(RS) &\subseteq \operatorname{SRG}(R)\operatorname{SRG}(S) \\ \operatorname{SRG}(R^{-1}) &= \operatorname{SRG}(R)^{-1}. \end{aligned}$$

For the precise meanings of these operations, and the requisite conditions on the systems R and S, we refer the reader to (Ryu et al., 2021).

2.5 SRGs of systems

The SRGs of LTI transfer functions are closely related to the Nyquist diagram, and the SRGs of static nonlinearities are closely related to the incremental circle. These connections are explored in detail in (Chaffey et al., 2021; Pates, 2021); below we recall the two main results. The h-convex hull is the regular convex hull with straight lines replaced by arcs with centre on the real axis – for a precise treatment, we refer the reader to (Huang et al., 2020).

Theorem 1. Let $g : L_2(\mathbb{C}) \to L_2(\mathbb{C})$ be linear and time invariant, with transfer function $G(s)$. Then $\operatorname{SRG}(g) \cap \mathbb{C}_{\operatorname{Im}\geq 0}$ is the h-convex hull of Nyquist $(G) \cap \mathbb{C}_{\operatorname{Im}\geq 0}$.

Theorem 2. Suppose $S : L_2 \to L_2$ is the operator given by a SISO static nonlinearity $s : \mathbb{R} \to \mathbb{R}$, such that for all $u_1, u_2 \in \mathbb{R}$, $y_i \in s(u_i)$,

$$\mu(u_1-u_2)^2 \leq (y_1-y_2)(u_1-u_2) \leq \lambda(u_1-u_2)^2. \quad (1)$$

Then the SRG of S is contained within the disc centred at $(\mu+\lambda)/2$ with radius $(\mu-\lambda)/2$.

Theorems 1 and 2, and the SRG interconnection rules, allow us to construct bounding SRGs for arbitrary interconnections of LTI and static nonlinear components. A simple example is illustrated in Figure 1.

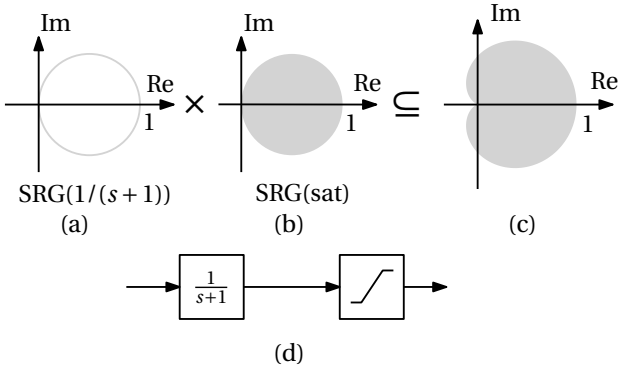


Fig. 1. Bounding SRG for the composition of a first order lag and saturation.

3. INCREMENTAL ROBUSTNESS AND SENSITIVITY

3.1 Stability and incremental gain

Given the negative feedback interconnection of Figure 2, incremental stability is guaranteed by the separation of the SRGs of P^{-1} and $-C$, and the distance between them is an incremental robustness margin, the reciprocal of which bounds the incremental gain of the feedback system. This is formalized in (Chaffey et al., 2021, Thm. 2); we recall the result here.

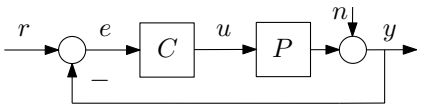


Fig. 2. Negative feedback control structure.

Let $\mathcal{H}$ be a class of operators. By $\bar{\mathcal{H}}$, we will denote a class of operators such that $\mathcal{H} \subseteq \bar{\mathcal{H}}$ and $\mathrm{SRG}(\bar{\mathcal{H}})$ satisfies the chord property: if $z_1, z_2 \in \mathrm{SRG}(\bar{\mathcal{H}})$, then $\vartheta z_1 + (1-\vartheta)z_2 \in \mathrm{SRG}(\bar{\mathcal{H}})$ for all $\vartheta \in [0,1]$.

Theorem 3. Consider the feedback interconnection shown in Figure 2 between any pair of operators $C \in \mathcal{C}$ and $P \in \mathcal{P}$, where $\mathcal{C}$ and $\mathcal{P}$ are classes of operators on L_2 with finite incremental gain. If, for all $\tau \in (0,1]$,

$$\mathrm{SRG}(\mathcal{C})^{-1} \cap -\tau\,\mathrm{SRG}(\bar{\mathcal{P}}) = \varnothing,$$

then the incremental L_2 gain from r to u is bounded by $1/r_m$, where r_m is the shortest distance between $\mathrm{SRG}(\mathcal{C}^{-1})$ and $-\mathrm{SRG}(\bar{\mathcal{P}})$.

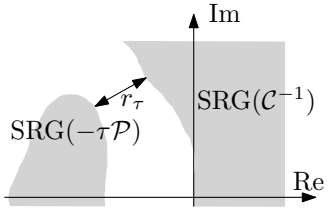


Fig. 3. Illustration of Theorem 3.

3.2 The sensitivity SRG

The operator $(I+PC)^{-1}$ maps r to e in the feedback system of Figure 2, and the operator $(I-PC(-I))^{-1}$ maps n to y. These two operators have the same SRG, which we denote by $\mathcal{S}$ – the *sensitivity SRG*.

Definition 4. The *peak incremental sensitivity* is the maximum incremental gain of the operator $(I+PC)^{-1}$.

The peak incremental sensitivity is equal to the maximum modulus of $\mathcal{S}$. The following theorem gives the peak incremental sensitivity an interpretation as a robustness margin.

Theorem 5. Let s_m be the shortest distance between $\mathrm{SRG}(PC)$ and the point -1. Then the peak incremental sensitivity is equal to $1/s_m$.

4. LOOP SHAPING

We demonstrate SRG loop shaping with two simple design examples for the control structure shown in Figure 2.

4.1 Shaping for stability and robustness

As first design example, we show how to use SRGs to ensure incremental stability of a closed loop system. Unlike traditional loop shaping, where the return ratio $L = PC$ is modified, we graphically shape the inverse of the feedback system, $(P+C^{-1})$, to improve the robustness of the closed loop. The use of SRGs makes the design close to classical Nyquist analysis, despite the nonlinearity of P.

Consider the system in Figure 4. C represents the controller, to be designed. Suppose that the process consists

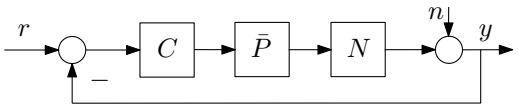


Fig. 4. Example control system. $\bar{P} = 1/(s(s+1))$, N is a nonlinear operator and C is the controller, to be designed. r is the reference input, n represents sensor noise.

of $\bar{P}$ with LTI dynamics $1/(s(s+1))$, and a nonlinear operator N, whose SRG is known to be bounded in the region illustrated in Figure 1 (c). We denote $C\bar{P}$ by L. The controller C is to be designed to stabilize the system and decrease the incremental gain.

To ensure stability, we require the SRGs of $L^{-1} = (C\bar{P})^{-1}$ and $-N$ to be separated, for all scalings of N between 0 and 1 (following Theorem 3). With $C_0 = 1$, the closed loop is unstable, as shown in Figure 5 (a). Shifting L^{-1} to the left, by designing C to give $L^{-1} = s(s+1)+1$, gives a stabilizing control. The controller reads $C_1 = s(s+1)/(1+s(s+1))$.

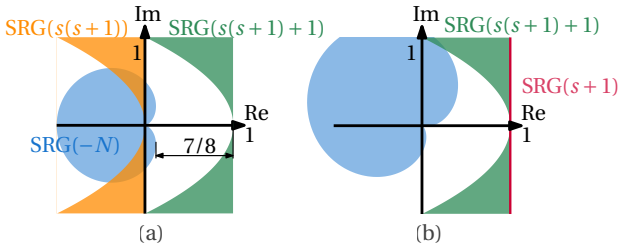


Fig. 5. (a) SRGs of $-N$, $\bar{P}^{-1}$ and $(\bar{P}C_1)^{-1}$. (b) SRGs of $(\bar{P}C_1)^{-1}$, $(\bar{P}C_2)^{-1}$ and a scaled and rotated nonlinearity, showing the improved robustness with C_2.

To improve robustness and reduce the incremental gain of the system, the separation of $\mathrm{SRG}(L^{-1})$ and $\mathrm{SRG}(-N)$ must be increased (again, following Theorem 3). For example, setting $L^{-1} = s+1$ ($C_2 = s$) gives good separation, and an incremental gain bound from r to u of $8/7 \approx 1.14$. As the incremental gain of N is bounded by 1 (the maximum modulus of its SRG), this value also bounds the incremental gain from r to y. The increased separation of the SRGs makes the system robust to uncertainties in the nonlinearity N, as illustrated in Figure 5 (b).

4.2 Shaping for performance

We now focus on graphical methods for improving performance, and explore how the sensitivity SRG can be shaped over particular sets of signals. We consider a new system, again of the form of Figure 2, with $C = 1/(ks+1)$, where k is a scalar to be designed, and P is a unit saturation. The SRGs of C and P are shown in Figure 1 (a) and (b).

Tracking performance and noise rejection are both characterized by the sensitivity SRG. Suppose that we would like this SRG to have a low modulus (corresponding to incremental gain) for signals with a bandwidth of $\omega_0 = 10$ rad/s and a maximum amplitude of 2. The aim is to limit the maximum amplification of $(I+PC)^{-1}$ over this range of signals.

A heuristic method is to maximize the distance between $\mathrm{SRG}(PC)$ and -1 over the frequency range $[-\omega_0, \omega_0]$

and amplitude range $[-2, 2]$, following Theorem 5. This corresponds to maximizing the minimum incremental gain of the inverse of the sensitivity operator over this range of signals.

SRG (PC) is bounded by the Minkowski product of SRG (P) and SRG (C). Plotting the SRG of the saturation P over the amplitude range $[-2, 2]$ gives the half-disc shown in Figure 6 (b). The SRG of C over $[-\omega_0, \omega_0]$ is described by

$$\left(\frac{1}{1+k^2\omega^2}, j\frac{-k\omega}{1+k^2\omega^2}\right)$$

for $\omega \in [-\omega_0, \omega_0]$. As a first design, we can set k so that the bandlimited SRG of C is half the circle (Figure 6 (a)) – this is achieved by setting $k = 0.01$. This gives the bound on SRG (PC) shown in Figure 6 (c). The minimum distance to the point -1 is $s_m = \sqrt{3}$.

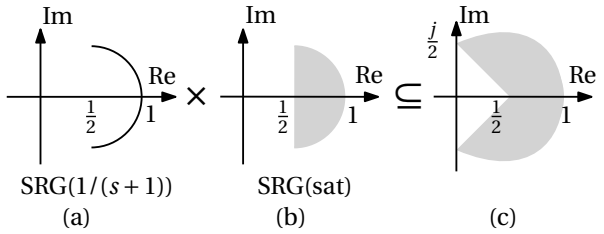


Fig. 6. Left: SRG of $1/(ks+1)$ over signals with bandwidth $[-1/k, 1/k]$. Right: SRG of saturation over signals with maximum amplitude 2.

This method is, however, only a heuristic. The saturation introduces higher harmonics, so the assumption that signals have a bounded spectrum is invalidated when the loop is closed. However, given the lowpass properties of the system, the approximation is reasonable. The higher order harmonics of the output of the saturation have low magnitude, and the unit lag has a lowpass behavior. The stability of the closed loop guarantees that these high frequencies are indeed attenuated by the feedback system. This assumption is similar to the lowpass assumption of describing function analysis Slotine and Li (1991). The method here differs from describing function analysis, however, in that arbitrary differences of bandlimited inputs are considered, not just pure sinusoids.

5. OTHER TYPES OF SYSTEMS

A significant advantage of the SRG is being able to place disparate system types on an equal footing. Like continuous time LTI systems, finite dimensional linear operators described by matrices lend themselves well to shaping. Pates (2021) has shown that the SRG of a matrix is equal to the numerical range of a closely related, transformed matrix. In the case of normal matrices, the SRG is the h-convex hull of the spectrum (Huang et al., 2020). These results pave the way for shaping a matrix's SRG by matrix multiplication and addition.

In cases where the analytic SRG is not available, the SRG can be sampled over the signals of interest, and loop shaping methods can then be applied using the sampled SRG. For example, Figure 7 shows a sampled SRG of the potassium conductance of the Hodgkin-Huxley model of a neuron (Hodgkin and Huxley, 1952).

REFERENCES

Bode, H.W. (1960). Feedback – the history of an idea. In *Proceedings of the Symposium on Active Networks and Feedback Systems*, Microwave Research Institute Symposia Series. Polytechnic Press, Brooklyn.

Chaffey, T. (2022). A rolled-off passivity theorem. *Systems & Control Letters*, 162.

Chaffey, T., Forni, F., and Sepulchre, R. (2021). Graphical nonlinear system analysis. *arXiv:2107.11272 [cs, eess, math]*.

Clegg, J.C. (1958). A nonlinear integrator for servomechanisms. *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry*, 77(1), 41–42. doi:10.1109/TAI.1958.6367399.

Hodgkin, A.L. and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4), 500–544. doi: 10.1113/jphysiol.1952.sp004764.

Huang, X., Ryu, E.K., and Yin, W. (2020). Scaled relative graph of normal matrices. *arXiv:2001.02061 [cs, math]*.

McFarlane, D. and Glover, K. (1992). A loop-shaping design procedure using H/sub infinity / synthesis. *IEEE Transactions on Automatic Control*, 37(6), 759–769. doi: 10.1109/9.256330.

Pates, R. (2021). The scaled relative graph of a linear operator. *arXiv:2106.05650 [math]*.

Ryu, E.K., Hannah, R., and Yin, W. (2021). Scaled relative graphs: Nonexpansive operators via 2D Euclidean geometry. *Mathematical Programming*. doi: 10.1007/s10107-021-01639-w.

Slotine, J.J.E. and Li, W. (1991). *Applied Nonlinear Control*. Prentice Hall, Englewood Cliffs, N.J.

Vinnicombe, G. (2000). *Uncertainty and Feedback: H_∞ Loop-Shaping and the ν-Gap Metric*. Imperial College Press.

Zames, G. (1981). Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26(2), 301–320. doi: 10.1109/TAC.1981.1102603.

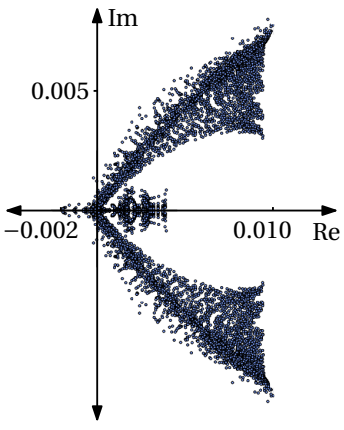


Fig. 7. Sampling of the SRG of a potassium conductance.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Evaluation Subspace Codes and Convolutional Codes

Joachim Rosenthal

Institute of Mathematics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich
(e-mail: rosenthal@math.uzh.ch)

Abstract: A constant dimension subspace code can be viewed geometrically as a subset of the Grassmann variety defined over a finite field.
There exist few algebraic constructions for constant dimension subspace codes. A major technique is the 'lifting technique' of a rank metric code with a good distance. For rank metric codes exist several good algebraic constructions. First and for most one should mention the technique of constructing Gabidulin codes which can be seen as the image of a linear space of linearized functions under an evaluation map. The technique of constructing Gabidulin codes naturally generalizes the construction of AG-codes such as Reed-Solomon codes and more general geometric Goppa codes.
In this talk we present a new idea on how one can construct excellent subspace codes by evaluating points on a rational curve in the Grassmannian.

Keywords: Subspace codes, rank metric codes, evaluation codes.

1. INTRODUCTION TO SUBSPACE CODES

Constant dimensional subspace codes appeared probably first in the seminal paper by Kötter and Kschischang (2008) on random network coding.

From a mathematical point of view we can view a constant dimensional subspace code simply as a subset of the finite Grassmann variety $\text{Grass}(k, \mathbb{F}_q^n)$ defined over some finite field $\mathbb{F}_q$.

One has a natural distance function on the Grassmannian. For this assume $\mathcal{U}, \mathcal{V} \subseteq \mathbb{F}_q^n$ are two subspaces. Then one defines their distance through:

$$d_S(\mathcal{U}, \mathcal{V}) := \dim(\mathcal{U} + \mathcal{V}) - \dim(\mathcal{U} \cap \mathcal{V}). \tag{1}$$

Once the distance for code elements is defined one gets a natural notion of distance for a subspace code by simply defining it as the minimal distance between two elements.

As in the classical theory of linear block codes it is a major design problem to come up with algebraic construction of constant dimension subspace codes whose distance is optimal or near optimal.

A survey on the construction of constant dimensional subspace codes is given in Horlemann-Trautmann and Rosenthal (2018).

As in the classical literature on coding theory one defines the distance $\text{dist}(\mathcal{C})$ of a code $\mathcal{C}$ as the minimal distance of two different code elements. The goal is once more the construction of subspace codes with large distance and many code elements.

⋆ Research supported by Swiss National Science Foundation under grant no. 188430

2. SOME KNOWN ALGEBRAIC CONSTRUCTIONS OF CONSTANT DIMENSION SUBSPACE CODES

Different from the theory of linear block codes there exist relatively few algebraic constructions of constant dimension subspace codes.

A major idea was already introduced by Kötter and Kschischang (2008) who showed that good rank metric codes give raise to good subspace codes by embedding the rank metric code into the 'thick open cell' of the Grassmann variety. The distance of the subspace code is then bounded by the distance of the underlying rank metric code.

In terms of matrices this lifting construction can be seen in the following way: Let

$$\{M_i \in \text{Mat}_{k\times m} \mid i = 1, \ldots, N\}$$

be a rank metric code defined over some finite field $\mathbb{F}_q$ and having N elements and minimum distance δ. Then the subspace code

$$\left\{\text{rowspace}_{\mathbb{F}_q}[I_k\ M_i] \in Grass(k, \mathbb{F}_q^{k+m}) \mid i = 1, \ldots, N\right\}$$

has also N elements and distance 2δ as one readily verifies.

Another major idea is to study group actions on the Grassmannian and to consider the orbit under this group actions. This then leads to the concept of 'Orbit codes' Trautmann et al. (2010) and the concept of 'cyclic orbit codes' Trautmann et al. (2013).

Beyond above construction techniques there exist several constructions techniques using geometric and combinatorial designs. The interested reader will find a survey in Horlemann-Trautmann and Rosenthal (2018).

It would certainly be desirable to have also construction techniques for constant dimension subspace codes which are based on some algebraic evaluation technique as this is done very successfully for Algebraic Geometric Goppa codes.

In this talk we will show how the associated Hermann Martin curve to a linear system (or a convolutional code) can be used to construct subspace codes via evaluation in some possible extension field. In the next section we will describe this technique. To our knowledge no similar constructions of subspace codes using evaluation maps is known.

3. THE ASSOCIATED HERMANN-MARTIN CURVE TO A LINEAR SYSTEM

It was an important contribution of Martin and Hermann (1978) that every linear system defines in a natural way a curve of genus zero in a Grassmann variety. One often calls the resulting curve the Hermann-Martin curve induced by the linear system. To make this concept a little more precise let $\mathbb{K}$ be an arbitrary field and consider a $k \times m$ transfer function $G(s)$.

Definition 1. Let $G(s)$ be a $k \times m$ transfer function and consider the map

$$h: \ \mathbb{K} \longrightarrow \mathrm{Grass}(k, \mathbb{K}^{k+m}), \ s \mapsto \mathrm{rowspace}_{\mathbb{K}}[I_k \ G(s)]. \quad (2)$$

Then h is called the *Hermann-Martin map* associated to the transfer function $G(s)$.

As this map is a rational map all the poles are removable. In order to see this in term of matrices consider a minimal left coprime factorization of the transfer function $G(s) = D^{-1}(s)N(s)$. Then h is equivalently described through:

$$h: \ \mathbb{K} \longrightarrow \mathrm{Grass}(k, \mathbb{K}^{p+m}), \ s \mapsto \mathrm{rowspace}_{\mathbb{K}}[D(s) \ N(s)]. \quad (3)$$

Note also that by the properties of a minimal left coprime factorization one has that $\mathrm{rowspace}_{\mathbb{K}}[D(\alpha) \ N(\alpha)]$ has full row rank for all elements α in the algebraic closure of $\mathbb{K}$ and the map can even be extended to the whole projective line $\mathbb{P}^1_{\mathbb{K}}$.

The identification by Martin and Hermann (1978) goes actually further as the McMillan degree of the transfer function corresponds to the degree of the Hermann-Martin curve and the observability indices of the linear system correspond to the Grothendick indices of an associated bundle over the projective line. (The pull back of the tautological bundle using the Hermann-Martin map). In the convolutional codes literature the observabilities indices are also often referred to as the Forney indices Forney, Jr. (1975) of the convolutional code and the concepts also naturally translate here.

The interested reader will find more material on these intriguing connections in Rosenthal (2005) and material on convolutional codes can be found in the recent survey article Lieb et al. (2021).

In the next section we will show how it is possible to construct excellent subspace codes startinng from a linear system (or convolutional code) and using evaluation of the the Hermann-Martin map.

4. A CONSTRUCTION OF SUBSPACE CODES USING THE HERMANN MARTIN MAP

In the sequel assume that the base field is the finite field $\mathbb{F}_q$. If $G(s)$ is a transfer function having the left coprime factorization $G(s) = D^{-1}(s)N(s)$ then we know from systems theory that $\mathrm{rowspace}_{\mathbb{K}}[D(\alpha) \ N(\alpha)]$ has full row rank for all elements α in any extension field of $\mathbb{F}_q$.

Based on this observation one can define a subspace code through:

$$\{\mathrm{rowspace}_{\mathbb{K}}[D(\alpha) \ N(\alpha)] \mid \alpha \in \mathbb{K}\}, \quad (4)$$

where $\mathbb{K} = \mathbb{F}_{q^k}$ is a finite extension field of $\mathbb{F}_q$.

Of course note that the resulting subspace code is a subspace code in the Grassmannian $\mathrm{Grass}(k, \mathbb{K}^n)$ defined over the extension field. It is also not clear how good the codes can be if one does such an evaluation.

In the sequel we will show how one can overcome the difficulties with the extension field and how it is possible to come up with excellent constant dimension subspace codes.

For simplicity we will consider only one input, one output transfer functions

$$G(s) := \frac{n(s)}{d(s)} \in \mathbb{F}_q(s).$$

Here $n(s), d(s)$ are simply elements of the polynomial ring $\mathbb{F}_q[s]$.

In order to present our main result we first have to make a definition:

Definition 2. A rational function $\frac{n(s)}{d(s)} \in \mathbb{F}_q(s)$ is called a permuation rational function over the extension field $\mathbb{K} = \mathbb{F}_{q^k}$ if the map

$$\varphi: \ \mathbb{K} \longrightarrow \mathbb{K}, \ s \mapsto \frac{n(s)}{d(s)} \quad (5)$$

describes a permutation of the extension field $\mathbb{K}$.

Note that if the denominator polynomial $d(s)$ is a constant then a permutation rational function is simply a permutation polynomial for $\mathbb{K} = \mathbb{F}_{q^k}$ this remark also shows that permutation rational functions are plentiful.

As it is well known one can identify elements of the extension field $\mathbb{K} = \mathbb{F}_{q^k}$ with the $\mathbb{F}_q$ algebra $\mathbb{F}_q[M]$ where M is a $k \times k$ matrix defined over $\mathbb{F}_q$ whose characteristic polynomial is irreducible.

A main result which hopefully justifies the outlined construction technique is then:

Theorem 3. Assume $\frac{n(s)}{d(s)} \in \mathbb{F}_q(s)$ is a permutation rational function over some extension field $\mathbb{K} = \mathbb{F}_{q^k}$ which is also non-proper, i.e. numerator degree is larger than the denominator degree. Identify elements of $\mathbb{K}$ with elements of the $\mathbb{F}_q$ algebra $\mathbb{F}_q[M]$. Then the evaluation map:

$$\mathbb{P}^1_{\mathbb{K}} \longrightarrow \mathrm{Grass}(k, \mathbb{K}^{2k}), \ \alpha \mapsto \mathrm{rowspace}_{\mathbb{K}}[d(\alpha) \, n(\alpha)]$$

defines a so called spread code. In particular this is a subspace code having maximal possible distance $2k$ and the number of elements is $q^k + 1$, the maximal possible cardinality.

Spread codes are somehow optimal subspace codes. The proof of the theorem follows from the way spread codes were constructed in Manganiello et al. (2008).

It will be a matter of future research to investigate if other evaluation maps can lead to subspace codes with optimal or near optimal distance.

Here one should in particular also look at the evaluation map based on multidimensional convolutional codes Weiner (1998).

REFERENCES

Forney, Jr., G.D. (1975). Minimal bases of rational vector spaces, with applications to multivariable linear systems. *SIAM J. Control*, 13(3), 493–520.

Horlemann-Trautmann, A. and Rosenthal, J. (2018). Constructions of constant dimension codes. In M. Greferath, M. Pavcevic, N. Silberstein, and M. Vazquez-Castro (eds.), *Network Coding and Subspace Design*, Signals and Communication Technology, 25–42. Springer Verlag.

Kötter, R. and Kschischang, F. (2008). Coding for errors and erasures in random network coding. *IEEE Transactions on Information Theory*, 54(8), 3579–3591. doi: 10.1109/TIT.2008.926449.

Lieb, J., Pinto, R., and Rosenthal, J. (2021). Convolutional codes. In J.S.P. Huffman C; Kim (ed.), *Concise Encyclopedia of Coding Theory*. CRC Press.

Manganiello, F., Gorla, E., and Rosenthal, J. (2008). Spread codes and spread decoding in network coding. In *Proceedings of the 2008 IEEE International Symposium on Information Theory*, 851–855. Toronto, Canada. doi: 10.1109/ISIT.2008.4595113.

Martin, C.F. and Hermann, R. (1978). Applications of algebraic geometry to system theory: The McMillan degree and Kronecker indices as topological and holomorphic invariants. *SIAM J. Control Optim.*, 16, 743–755.

Rosenthal, J. (2005). The Hermann-Martin curve. In *New directions and applications in control theory*, volume 321 of *Lecture Notes in Control and Inform. Sci.*, 353–365. Springer, Berlin.

Trautmann, A.L., Manganiello, F., Braun, M., and Rosenthal, J. (2013). Cyclic orbit codes. *IEEE Trans. Inform. Theory*, 59(11), 7386–7404.

Trautmann, A.L., Manganiello, F., and Rosenthal, J. (2010). Orbit codes - a new concept in the area of network coding. In *Information Theory Workshop (ITW), 2010 IEEE*, 1 –4. Dublin, Ireland. doi: 10.1109/CIG.2010.5592788.

Weiner, P. (1998). *Multidimensional Convolutional Codes*. Ph.D. thesis, University of Notre Dame.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A Finite Lyapunov Matrix-Based Stability Criterion via Piecewise Linear Approximation ⋆

Irina V. Alexandrova *

* *St. Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg, 199034, Russia (e-mail: i.v.aleksandrova@spbu.ru).*

Abstract: Recently, a number of Lyapunov matrix-based necessary and sufficient stability tests which require a finite set of operations to be verified were presented for linear time-invariant time delay systems, see Egorov et al. (2017), Gomez et al. (2019) and Bajodek et al. (2022). Motivated by those works, in this contribution we revisit the early paper Medvedeva & Zhabko (2015) and develop the idea to construct a necessary and sufficient finite stability test for a single-delay system as well. The approach relies on a simple piecewise linear approximation of the arguments of Lyapunov–Krasovskii functionals based on the Lyapunov matrix, and shows its competitiveness at least in case of small delays.

Keywords: time delay systems, linear systems, exponential stability, Lyapunov–Krasovskii functionals, Lyapunov matrices

The Lyapunov–Krasovskii functionals with prescribed derivative based on the so-called Lyapunov matrix are known to deliver the necessary and sufficient stability conditions for linear time-invariant time delay systems (Kharitonov & Zhabko, 2003). However, verification of their positive definiteness required for the stability test is challenging. Recently, a number of finite necessary and sufficient tests where such verification is reduced to positive definiteness analysis of a certain block matrix based on the Lyapunov matrix values have been appeared. In Egorov et al. (2017); Egorov & Mondié (2014); Gomez et al. (2019), this was achieved by approximating the functionals arguments by specific piecewise continuous functions based on the fundamental matrix of the system. The approach of Bajodek et al. (2022) employs a different type of approximation based on Legendre polynomials. The benefit of the first approach is that the resulting block matrix has very nice structure. Indeed, it consists solely of the Lyapunov matrix values at different points of the delay interval even for systems with multiple delays. In the second approach, the Lyapunov matrix-based integrals are involved in the matrix. However, its dimension is much smaller, which is a significant advantage.

It is important to note that both approaches rely on the key ideas of early works (Medvedeva & Zhabko, 2013, 2015) which made the construction of finite necessary and sufficient Lyapunov matrix-based stability tests possible. Those ideas include introducing a specific set of functions satisfying the condition $\|\varphi(\theta)\| \leqslant \|\varphi(0)\|$ for $\theta \in [-h, 0]$ in context of positive definiteness test; showing that the approximation error (i.e. the difference between the exact and the approximated functionals) tends towards zero when the discretization is refined; estimating the exact functional on the solution from the mentioned set in case of instability. The numerical scheme proposed in Medvedeva & Zhabko (2013, 2015) employed simply piecewise linear or piecewise cubic approximations of the functionals arguments. Here, being motivated by Egorov et al. (2017); Gomez et al. (2019) and Bajodek et al. (2022), we give a new perspective on the piecewise linear approximation scheme of Medvedeva & Zhabko (2015). The goal is to find a compromise between the resulting block matrix dimension and its simplicity as well as whole computational performance of the approach.

Note that our approach is different from those of Gu (1997) where the functionals kernels rather than arguments are approximated by piecewise linear matrix functions.

Notation: Assume that initial functions belong to the space $PC\big([-h, 0], \mathbb{R}^n\big)$ of piecewise continuous $\mathbb{R}^n$-valued functions defined on $[-h, 0]$ which is supplied with the norm $\|\varphi\|_h = \sup_{\theta\in[-h,0]} \|\varphi(\theta)\|$; $C^2\big([-h, 0], \mathbb{R}^n\big)$ stands for the space of twice continuously differentiable vector functions; $\Re(s)$ denotes the real part of a complex value s; $\lambda_{\min}(W)$ is the smallest eigenvalue of a matrix W; notation $k = \overline{n_1, n_2}$, where $n_1, n_2 \in \mathbb{Z}$, $n_1 < n_2$, means that k is an integer between n_1 and n_2; $\lceil\cdot\rceil$ denotes the ceiling function; $\mathrm{vec}(X)$ means a vectorization of the matrix X; $A \otimes B$ stands for the Kronecker product, namely,

$$A \otimes B \stackrel{\text{def}}{=} \begin{pmatrix} b_{11}A & b_{21}A & \dots & b_{n1}A \\ b_{12}A & b_{22}A & \dots & b_{n2}A \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n}A & b_{2n}A & \dots & b_{nn}A \end{pmatrix},$$

where $B = \big\{b_{ij}\big\}_{i,j=1}^{n}$.

In this work, we analyze the exponential stability of a linear time delay system of the form

$$\dot{x}(t) = A_0 x(t) + A_1 x(t - h), \quad t \geqslant 0. \tag{1}$$

⋆ The work was supported by the Russian President Program for promotion of young researchers, Project MK-2301.2022.1.1.

Here, A_0, A_1 are constant $n \times n$ matrices, and $h \geqslant 0$ is a constant delay. Given a positive definite matrix W, consider a functional with the derivative prescribed along the solutions of system (1) as

$$\frac{\mathrm{d}v_0(x_t)}{\mathrm{d}t} = -x^T(t)Wx(t), \quad t \geqslant 0.$$

This functional is given by (Huang, 1989)

$$v_0(\varphi) = \varphi^T(0)U(0)\varphi(0) + 2\varphi^T(0)\int_{-h}^{0} U^T(h+\theta)A_1\varphi(\theta)\mathrm{d}\theta + \int_{-h}^{0}\int_{-h}^{0} \varphi^T(\theta_1)A_1^T U(\theta_1-\theta_2)A_1\varphi(\theta_2)\mathrm{d}\theta_2\mathrm{d}\theta_1.$$

Here, $U(\tau)$, $\tau \in [-h,h]$, is called the Lyapunov matrix of system (1) associated with W. This matrix can be computed as a matrix exponential provided that the so-called Lyapunov condition holds, see Kharitonov (2013) for the details. Now, introduce the set

$$\mathcal{S} = \Big\{\varphi \in C^2([-h,0],\mathbb{R}^n) \;\Big|\; \|\varphi\|_h = \|\varphi(0)\| = 1, \; \|\varphi^{(l)}\|_h \leqslant K^l, \; l = 1,2\Big\},$$

where $K = \|A_0\| + \|A_1\|$, and $\varphi^{(l)}$ means the l-th derivative of φ. It is shown in Medvedeva & Zhabko (2013, 2015) that system (1) is exponentially stable, if and only if there exist functional v_0 and a constant $\mu > 0$ such that

$$v_0(\varphi) \geqslant \mu\|\varphi(0)\|^2, \quad \varphi \in \mathcal{S}.$$

In other words, positive definiteness of the functional may be assessed on the set $\mathcal{S}$ only for the stability test of system (1). Moreover, if system (1) is unstable, then there exists a function $\varphi \in \mathcal{S}$ such that

$$v_0(\varphi) \leqslant -a_0 \stackrel{\text{def}}{=} -\lambda_{\min}(W)/(4\hat{\alpha}),$$

where $\hat{\alpha}$ is such that $\Re(s) \leqslant \hat{\alpha}$ with s being an eigenvalue of an unstable system (1). In particular, $\hat{\alpha} = K$ may be taken as a rough bound. The last-mentioned claim can be found in the implicit form in Alexandrova & Zhabko (2019).

Below, we revisit a piecewise linear discretization scheme of a function φ in functional v_0 proposed in Medvedeva & Zhabko (2015). First, discretize an interval $[-h,0]$ with the points $\theta_j = -j\Delta$, $j = \overline{0,N}$, and consider a piecewise linear approximation of a function φ given by

$$\varphi(s+\theta_j) = l_N(s+\theta_j) + \eta_N(s+\theta_j), \;\; s \in [-\Delta, 0], \quad (2)$$

$j = \overline{0,N-1}$, where

$$l_N(s+\theta_j) = \varphi(\theta_j)\left(1+\frac{s}{\Delta}\right) - \varphi(\theta_{j+1})\frac{s}{\Delta}. \quad (3)$$

Here, $l_N(\theta)$ and $\eta_N(\theta)$, $\theta \in [-h,0]$, stand for the approximation itself and the approximation error, respectively. Now, we substitute the approximation (2)–(3) into the functional $v_0(\varphi)$ and arrive at the expression

$$v_0(\varphi) = v_0(l_N) + \Psi_N.$$

It turns out that the approximated functional $v_0(l_N)$ represents a quadratic form with respect to the vector

$$\widehat{\varphi} = \begin{pmatrix} \varphi(0) \\ \varphi(\theta_1) \\ \vdots \\ \varphi(\theta_N) \end{pmatrix}.$$

That is,

$$v_0(l_N) = \widehat{\varphi}^T \mathcal{L}_N \widehat{\varphi},$$

where $\mathcal{L}_N = \left\{\mathcal{L}^{ij}\right\}_{i,j=0}^{N}$ consists of the blocks $\mathcal{L}^{ij}$ of dimensions $n \times n$. Further, the piecewise linear approximation error admits a bound of the form

$$\|\eta_N(s+\theta_j)\| \leqslant \frac{1}{2}K^2(-s)(s+\Delta), \quad s \in [-\Delta, 0],$$

for all $j = \overline{0,N-1}$. Hence, we derive that

$$|\Psi_N| \leqslant \delta_N \stackrel{\text{def}}{=} \frac{c_1}{N^2} + \frac{c_2}{N^4},$$

where

$$c_1 = \frac{1}{6}K^2h^3(M_1 + hM_2), \quad c_2 = \frac{1}{144}M_2K^4h^6,$$
$$M_1 = \max_{\theta\in[0,h]}\|U^T(\theta)A_1\|, \;\; M_2 = \max_{\theta\in[-h,h]}\|A_1^T U(\theta)A_1\|.$$

Clearly, the term Ψ_N approaches zero when the value of N tends to infinity. Following Gomez et al. (2019) and Bajodek et al. (2022), we introduce the next lemma which plays a key role in the presentation of the result in the form a finite criterion.

Lemma 1. Given $\varepsilon > 0$, if

$$N \geqslant \sqrt{\frac{c_1 + \sqrt{c_1^2 + 4\varepsilon c_2}}{2\varepsilon}},$$

then $|\Psi_N| \leqslant \varepsilon$.

Finally, defining the value

$$N^\star = \left\lceil \sqrt{\frac{c_1 + \sqrt{c_1^2 + 4a_0c_2}}{2a_0}} \right\rceil,$$

we are ready present the following stability criterion.

Theorem 2. System (1) is exponentially stable, if and only if the Lyapunov condition holds and

$$\min_{\substack{\|\zeta_0\|=1,\\ \|\zeta_j\|\leqslant 1,\, j=\overline{1,N^\star}}} \zeta^T\mathcal{L}_{N^\star}\zeta > 0. \quad (4)$$

Here, $\zeta = \left(\zeta_0^T, \zeta_1^T, \ldots, \zeta_{N^\star}^T\right)^T \in \mathbb{R}^{n(N^\star+1)}$.

Note that instead of (4) positive semi-definiteness of matrix $\mathcal{L}_{N^\star+1}$ may be verified. Below, we present expressions for the blocks $\mathcal{L}^{ij}$ of the matrix $\mathcal{L}_N$ for completeness:

$$\mathcal{L}^{00} = U(0) + 2\int_{-\Delta}^{0} U^T(s+N\Delta)\left(1+\frac{s}{\Delta}\right)\mathrm{d}sA_1 + A_1^T\int_{-\Delta}^{0}\int_{-\Delta}^{0}\left(1+\frac{s_1}{\Delta}\right)\left(1+\frac{s_2}{\Delta}\right) \times U(s_1-s_2)\mathrm{d}s_2\mathrm{d}s_1A_1,$$

$$\mathcal{L}^{k,k+l} = A_1^T\int_{-\Delta}^{0}\int_{-\Delta}^{0}\Big\{\Big[\Big(1+\frac{s_1}{\Delta}\Big)\Big(1+\frac{s_2}{\Delta}\Big) + \frac{s_1s_2}{\Delta^2}\Big] \times U(s_1-s_2+l\Delta) + \Big(1+\frac{s_1}{\Delta}\Big)\Big(-\frac{s_2}{\Delta}\Big) \times U(s_1-s_2+(l-1)\Delta) + \Big(-\frac{s_1}{\Delta}\Big)\Big(1+\frac{s_2}{\Delta}\Big) \times U(s_1-s_2+(l+1)\Delta)\Big\}\mathrm{d}s_2\mathrm{d}s_1A_1,$$

$$\mathcal{L}^{NN} = A_1^T\int_{-\Delta}^{0}\int_{-\Delta}^{0}\frac{s_1s_2}{\Delta^2}U(s_1-s_2)\mathrm{d}s_2\mathrm{d}s_1A_1,$$

$$\begin{aligned}
\mathcal{L}^{0k} &= \int_{-\Delta}^{0} \Big[U^T(s+(N-k)\Delta)\left(1+\frac{s}{\Delta}\right) \\
&\quad + U^T(s+(N-k+1)\Delta)\left(-\frac{s}{\Delta}\right)\Big] \mathrm{d}s A_1 \\
&\quad + A_1^T \int_{-\Delta}^{0}\int_{-\Delta}^{0} \left(1+\frac{s_1}{\Delta}\right) \\
&\quad \times \Big[\left(1+\frac{s_2}{\Delta}\right) U(s_1-s_2+k\Delta) \\
&\quad + \left(-\frac{s_2}{\Delta}\right) U(s_1-s_2+(k-1)\Delta)\Big] \mathrm{d}s_2 \mathrm{d}s_1 A_1, \\
\mathcal{L}^{0N} &= \int_{-\Delta}^{0} U^T(s+\Delta)\left(-\frac{s}{\Delta}\right) \mathrm{d}s A_1 \\
&\quad + A_1^T \int_{-\Delta}^{0}\int_{-\Delta}^{0} \left(1+\frac{s_1}{\Delta}\right)\left(-\frac{s_2}{\Delta}\right) \\
&\quad \times U(s_1-s_2+(N-1)\Delta)\mathrm{d}s_2 \mathrm{d}s_1 A_1, \\
\mathcal{L}^{kN} &= A_1^T \int_{-\Delta}^{0}\int_{-\Delta}^{0} \Big[\left(1+\frac{s_1}{\Delta}\right) \\
&\quad \times U(s_1-s_2+(N-k-1)\Delta) + \left(-\frac{s_1}{\Delta}\right) \\
&\quad \times U(s_1-s_2+(N-k)\Delta)\Big] \left(-\frac{s_2}{\Delta}\right) \mathrm{d}s_2 \mathrm{d}s_1 A_1.
\end{aligned}$$

Here, $k = \overline{1, N-1}$, $l = \overline{0, N-k-1}$, and $\mathcal{L}^{jk} = \mathcal{L}^{kj\,T}$ for other indices.

Despite the fact that the matrix $\mathcal{L}_N$ is determined by the integrals of the Lyapunov matrix, multiplied possibly by polynomials, we claim that all those integrals may be computed explicitly in a vectorized form, without performing the operation of integration in fact, provided that $\det(L) \neq 0$, where

$$L = \begin{pmatrix} I \otimes A_0 & I \otimes A_1 \\ -A_1^T \otimes I & -A_0^T \otimes I \end{pmatrix}.$$

For instance, the terms

$$J_l = \int_{-\Delta}^{0} U(s+l\Delta)\mathrm{d}s, \quad l = \overline{1, N},$$

may be computed from

$$\begin{pmatrix} a_l \\ a^\star_{N-l+1} \end{pmatrix} = L^{-1} \begin{pmatrix} u(l\Delta) - u((l-1)\Delta) \\ u^\star((N-l)\Delta) - u^\star((N-l+1)\Delta) \end{pmatrix},$$

where

$$u(\tau) = \mathrm{vec}(U(\tau)), \quad u^\star(\tau) = \mathrm{vec}(U^T(\tau)),$$
$$a_l = \mathrm{vec}(J_l), \quad a_l^\star = \mathrm{vec}(J_l^T), \quad l = \overline{1, N}.$$

This fact allows us to improve computational performance of the approach significantly, comparing to the experiments made in Medvedeva & Zhabko (2015).

Finally, we suggest to use our approach in a combination with the necessary stability conditions of Egorov & Mondié (2014), since the values $U(k\Delta)$, $k = \overline{0, N}$, involved in their matrix are calculated during the computation of $\mathcal{L}_N$ anyway.

REFERENCES

I.V. Alexandrova, & A.P. Zhabko. Stability of neutral type delay systems: A joint Lyapunov–Krasovskii and Razumikhin approach. *Automatica*, 106, 83–90, 2019.

M. Bajodek, F. Gouaisbaut, A. Seuret. Necessary and sufficient stability condition for time-delay systems arising from Legendre approximation. Submitted to *IEEE Transactions on Automatic Control*, 2022. hal-03435028

A.V. Egorov, C. Cuvas, S. Mondié. Necessary and sufficient stability conditions for linear systems with pointwise and distributed delays. *Automatica*, 80, 218–224, 2017.

A.V. Egorov, & S. Mondié. Necessary stability conditions for linear delay systems. *Automatica*, 50, 3204–3208, 2014.

M. Gomez, A.V. Egorov, & S. Mondié. Lyapunov matrix based necessary and sufficient stability condition by finite number of mathematical operations for retarded type systems. *Automatica*, 108,108475, 2019.

K. Gu. Discretized LMI set in the stability problem of linear uncertain time delay systems. *International Journal of Control*, 68, 923–934, 1997.

W. Huang. Generalization of Liapunov's theorem in a linear delay system. *J. of Mathematical Analysis and Applications*, 142, 83–94, 1989.

V.L. Kharitonov. Time-Delay Systems: Lyapunov Functionals and Matrices. Basel: Birkhäuser, 2013.

V.L. Kharitonov, & A.P. Zhabko. Lyapunov – Krasovskii approach to the robust stability analysis of time-delay systems. *Automatica*, 39, 15–20, 2003.

I.V. Medvedeva, & A.P. Zhabko. Constructive method of linear systems with delay stability analysis. In *Proceedings of 11th IFAC Workshop on Time-Delay Systems*, Grenoble, France, 1–6, 2013.

I.V. Medvedeva, & A.P. Zhabko. Synthesis of Razumikhin and Lyapunov–Krasovskii approaches to stability analysis of time-delay systems. *Automatica*, 51, 372–377, 2015.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Signature-based models in finance: theory and calibration ⋆

Christa Cuchiero * **Guido Gazzani** ** **Sara Svaluto-Ferro** ***

* *Vienna University, Department of Statistics and Operations Research, Data Science @ Uni Vienna, 1090 Wien, Austria, (e-mail: christa.cuchiero@univie.ac.at).*
** *Vienna University, Department of Statistics and Operations Research, 1090 Wien, Austria, (e-mail: guido.gazzani@univie.ac.at).*
*** *University of Verona, Department of Economics, 37129 Verona, Italy, (e-mail: sara.svalutoferro@univr.it).*

Abstract: Signature methods represent a non-parametric way for extracting characteristic features from time series data which is essential in machine learning tasks. This explains why these techniques become more and more popular in econometrics and mathematical finance. Indeed, signature based approaches allow for data-driven and thus more robust model selection mechanisms, while first principles like no arbitrage can still be easily guaranteed.
Here we focus on financial models whose dynamics are described by linear functions of the (time-extended) signature of a primary underlying process, which can range from a (market-inferred) Brownian motion to a general multidimensional tractable stochastic process. The framework is universal in the sense that any classical model can be approximated arbitrarily well and that the model characteristics can be learned from all sources of available data by simple methods. In view of option pricing and calibration, key quantities that need to be computed in these models are the expected value or Fourier Laplace transform of the signature of the primary underlying process. Surprisingly this can be achieved via techniques from affine and polynomial processes. These formulas can then be used in the calibration procedure to option prices, while calibration to time series data just reduces to a simple regression.

Keywords: signature methods, calibration of financial models, Monte Carlo methods, linear infinite dimensional systems, affine and polynomial processes
MSC (2010) Classification: 91B70, 62P05, 65C20.

1. INTRODUCTION

In the past few years data driven models have successfully entered the area of stochastic modeling and mathematical finance. The paradigm of calibrating a few well interpretable parameters has changed to learning the model's characteristics as a whole, thereby exploiting all available sources of data. Thus highly parametric and over-parametrized models methods have gained more and more importance. On the one hand side this has opened the door to robust and more data-driven model selection mechanisms, while on the other hand model classes still have to be chosen in a way to guarantee first principles from finance like "no arbitrage". Relying on different universal approximation theorems then leads to different well-suited universal classes of dynamic processes that can serve both purposes.

One class of such financial models are so-called *neural stochastic differential equations* (SDEs) which are defined as Itô-diffusions where the drift and the volatility function are parameterized via neural networks (see e.g. Gierjatowicz et al. (2020); Cuchiero et al. (2020); Cohen et al. (2021)). Another class of models, considered in Perez-Arribas et al. (2020) and inspiring the current work, are so-called *Sig-SDEs*. These are again Itô-diffusions, however in this case the characteristics are linear functions of the *signature* (more precisely introduced below) of some driving Brownian motion and time.

We consider here a related approach, where the asset price model itself is parameterized as a linear function of the signature of a primary underlying process. This underlying process can either be a classical driving signal, e.g. a Brownian motion, but also a more general tractable stochastic model describing well observable quantities.

Before going into the details of the current model framework, let us first explain the mathematical significance of *signature*, a notion which goes back to Chen (1977, 1957) and plays a particular important role in the context of rough path theory initiated by Lyons (1998). Indeed, the signature of an $\mathbb{R}^d$-valued path serves as linear regression basis for continuous path functionals, since

- it is *point-separating*, as long as the path contains one strictly monotone component (which can always be achieved by adding time), as it then uniquely determines the underlying path;

⋆ The authors gratefully acknowledge financial support through grant Y 1235 of the FWF START-program.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

- linear functions on the signature form an algebra that contains 1. More precisely, every polynomial on the signature may be realized as a linear function via the so-called shuffle product.

The Stone-Weierstrass theorem therefore yields a universal approximation theorem (UAT), telling that continuous (with respect to a certain variation distance) path functionals on compact sets can be uniformly approximated by a linear function of the time extended signature. Therefore signature-based methods provide a non-parametric way to extract characteristic features (linearly) from time series data, which is essential in machine learning tasks in finance. This explains why these techniques become more and more popular in econometrics and mathematical finance, see e.g., Buehler et al. (2020); Perez-Arribas et al. (2020); Lyons et al. (2020); Ni et al. (2021); Bayer et al. (2021); Akyildirim et al. (2022) and the references therein.

We consider here signature-based methods with the goal to provide a data-driven, universal, tractable and easy to calibrate model for a set of traded assets $S = (S^1, \dots, S^m)$. To achieve this the main ingredient is a primary underlying process $(\widehat{X}_t)_{t\geq 0} = (t, X_t^1, \dots, X_t^d)_{t\geq 0}$ with $d \leq m$, where X is a continuous Itô-semimartingale. We suppose here that time-series data of $\widehat{X}$ is available and that its signature denoted by $\widehat{\mathbb{X}}$ serves a linear regression basis for S.

2. NOTATION AND PRELIMINARIES ON SIGNATURE

We shall now introduce the most essential concepts in order to rigorously define signature in the current context. In particular, we shall use the following notation:

- The signature takes values in the extended tensor algebra $T((\mathbb{R}^d))$ given by
$$T((\mathbb{R}^d)) := \{(a_0, \dots, a_n, \dots) \,|\, n \geq 0,\, a_n \in (\mathbb{R}^d)^{\otimes n}\}.$$
Elements of $T((\mathbb{R}^d))$ are denoted in bold face, e.g. $\mathbf{a} = (a_0, a_1, \dots, a_n, \dots)$.
- Let $I = (i_1, \dots, i_n)$ be a multi-index with entries in $\{1, \dots, d\}$ and denote by $e_I = e_{i_1} \otimes \cdots \otimes e_{i_n}$ the basis elements of $(\mathbb{R}^d)^{\otimes n}$.
- We write $\langle e_I, \mathbf{a}\rangle$ to extract the I^{th} component from a_n. More generally we often write $\mathbf{u}(\mathbf{x}) = \langle \mathbf{u}, \mathbf{x}\rangle$ if $\sum_I |u_I x_I| < \infty$ and call this linear maps in $\mathbf{x}$ (on their domain of convergence).

The coordinate signature indexed by a multi-index $I = (i_1, \dots, i_n)$ of an $\mathbb{R}^d$-valued semimartingale $\widehat{X}$ is defined via iterated Stratonovic integrals (denoted by $\circ$)
$$\langle e_I, \widehat{\mathbb{X}}_T\rangle := \int_{0<t_1<\cdots<t_n<T} \circ d\widehat{X}_{t_1}^{i_1} \cdots \circ\, d\widehat{X}_{t_n}^{i_n}.$$
Hence, $\widehat{\mathbb{X}}_T = 1 + \sum_{n=1}^{\infty} \sum_{|I|=n} \langle e_I, \widehat{\mathbb{X}}_T\rangle e_I \in T((\mathbb{R}^d))$.

3. THE MODEL AND ITS PROPERTIES

Let us now describe the precise modeling framework. The traded assets $(S^1, \dots, S^m)$ are modeled via $S_n(\ell) = (S_n^1(\ell), \dots, S_n^m(\ell))$ where
$$S_n^j(\ell^j)_t := \ell(\mathbb{X}_t) = \ell_0^j + \sum_{0<|I|\leq n} \ell_I^j \langle e_I, \widehat{\mathbb{X}}_t\rangle, \quad \text{(Sig-model)}$$
with

- $\widehat{\mathbb{X}}$ the signature of $\widehat{X}$,
- $n \in \mathbb{N}$ is the degree of truncation,
- $\ell_0^j, \ell_I^j \in \mathbb{R}$ are the deterministic coefficients of the linear map ℓ to be found from data.

For notational simplicity we shall in the sequel set $m = 1$.

Note that since it is possible to express (Sig-model) also in terms of stochastic integrals, the class of Sig-SDEs considered in Perez-Arribas et al. (2020) can be embedded in our framework by choosing a one-dimensional Brownian motion as primary underlying process.

The attractiveness of this model class described by (Sig-model) arises from several important features that we summarize in the sequel:

Universality: Any classical model driven by Brownian motion can be arbitrarily well approximated. This is again a consequence of the Stone-Weierstrass theorem because the solution map of a stochastic differential equation is a continuous (with respect to a certain variation distance) map of the signature of the driving signal.

No arbitrage: The model can also be expressed in terms of stochastic integrals with respect to local martingales, from which conditions for no-arbitrage can be easily deduced.

Tractable option pricing formulas: By relying on the above UAT and in turn on approximations via so-called sig-payoffs of the form $\langle e_J, \widehat{\mathbb{S}}_T(\ell)\rangle$ (see also Lyons et al. (2020)), (approximate) option pricing reduces to the computation of the expected signature of $\widehat{X}$. Thus the question is for which primary underlying processes $\mathbb{E}_{\mathbb{Q}}[\widehat{\mathbb{X}}_T]$ can be easily computed. This is the case for highly generic processes of the form
$$d\widehat{X}_t = \mathbf{b}(\widehat{\mathbb{X}}_t)dt + \sqrt{\mathbf{a}(\widehat{\mathbb{X}}_t)}dB_t, \qquad (1)$$
where $\mathbf{b}$ and $\mathbf{a}$ are linear maps. Indeed, as shown in Cuchiero et al. (2022) these processes can be seen as projections of extended tensor algebra valued *affine and polynomial process* (see Duffie et al. (2003); Cuchiero et al. (2012)), which implies that the expected signature can be computed by solving a linear ODE. This ODE is usually infinite dimensional, but if $\widehat{X}$ is itself a polynomial process it becomes finite dimensional. Analogously we can rely on affine technology, i.e. solving (infinite dimensional) Riccati equations to obtain the Fourier-Laplace transform of the marginals of $\widehat{\mathbb{X}}$.

Note that similarly to polynomial approximations the approximation of vanilla call and put option via sig-payoffs is not straightforward. Nevertheless sig-payoffs can be used for variance reduction techniques and are interesting in their own right as certain path dependent options like Asian forwards fall into this class.

Calibration to time series data: The tractability of the model class becomes particularly clear in view of calibration tasks. Indeed, when the goal is to calibrate to times series data of market prices $(S_{t_i}^M)_{i=1}^N$, this task reduces to a simple linear regression.

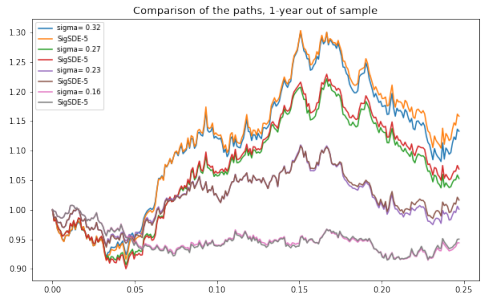


Fig. 1. Out of sample comparison between trajectories generated from a Black-Scholes model and the calibrated Sig-model.

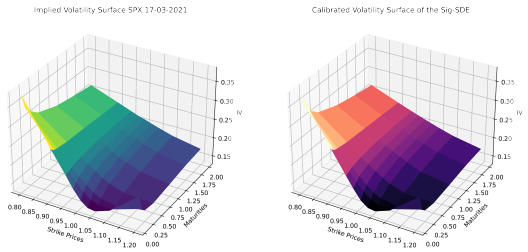


Fig. 2. Comparison of the market volatility surface (left) and the calibrated Sig-model volatility surface (right).

Calibration to options: When calibrating to the market's volatility surface, we exploit the linearity of the model by precomputing Monte-Carlo samples of $\widehat{\mathbb{X}}$ and then performing a standard optimizations to find the parameters of the linear map ℓ. By initializing the parameters of ℓ appropriately this optimization task actually becomes a convex problem, which makes it particularly tractable. On simulated and real market data (S&P 500 index) we show that a full calibration to the volatility surface, in particular when using time dependent parameters, is highly accurate and very fast.

Note that our calibration method differs from the approach proposed in Perez Arribas (2020) since we do not rely on sig-payoffs as a good approximation of vanilla call and put option via sig-payoffs uniformly over the range of possible models can hardly be achieved.

Subsequently we shall illustrate some of these model features in more detail.

3.1 Calibration to time series data

With regard to calibration to time-series data, the goal is to match N market prices $(S^M_{t_1}, \dots, S^M_{t_N})$. Due to our assumption that time series data of the primary underlying process $\widehat{X}_{t_1}, \dots, \widehat{X}_{t_N}$ (e.g. market inferred Brownian motion) is available, we can compute the path of its signature $\widehat{\mathbb{X}}$. This then serves as linear regression basis to find ℓ by matching the prices, i.e.

$$\text{argmin}_\ell \sum_{i=1}^{N} \left(\ell_0 + \sum_{0\le|I|\le n} \ell_I \langle e_I, \widehat{\mathbb{X}}_{t_i} \rangle - S^M_{t_i} \right)^2$$

Since the dimension of ℓ is typically high, introducing a regularization (Lasso, Ridge) is necessary. In the Figure 3.1 below we illustrate by means of a 4 dimensional Black-Scholes market how well out-of-sample trajectories can be learned. There we us as primary underlying process time extended (market inferred) Brownian motion up to order 5 and regress on its signature.

3.2 Calibration to option data

When calibrating to option data the goal is to match N option prices $(\pi^1, \dots, \pi^N)$ corresponding to European payoffs $F_i(S_{T_i})$ for $i = 1, \dots, N$. Typically we calibrate to call and put options with different strikes and maturities, whose prices are expressed in terms of implied volatility. To achieve this we start by computing M Monte-Carlo samples of $\widehat{\mathbb{X}}^j_{T_1}, \dots, \widehat{\mathbb{X}}^j_{T_N}$ for $j = 1, \dots M$ (under a pricing measure $\mathbb{Q}$). The calibration can then be formalized via

$$\text{argmin}_\ell \sum_{i=1}^{N} w^i \left(\frac{1}{M} \sum_{j=1}^{M} F_i(\ell(\widehat{\mathbb{X}}^j_{T_i})) - \pi_i \right)^2,$$

where w^i are weights, e.g. vega-weights known to match implied volatility well. The advantages of the Sig-model class is that all Monte-Carlo samples can be easily precomputed and re-used, so that the calibration reduces to a simple optimization task without any Monte-Carlo simulation in an optimization step. A further nice feature is that for parameters ℓ such that $\frac{1}{M}\sum_{j=1}^{M} F_i(\ell(\widehat{\mathbb{X}}^j_{T_i})) \ge \pi_i$ the optimization is convex for convex payoffs. This means that for starting values in this range and small learning rates in appropriate gradient descent methods the algorithms are likely to converge to the true minimum.

In Figure 3.2 we illustrate the calibration to an S&P 500 volatility surface (from 17-03-21) by using as primary underlying process a two dimensional time extended Brownian motion. We consider here an extension of the model with time-dependent parameters and achieve a nearly perfect fit. Indeed, with 13 parameters per maturity the absolute error is in the range of 0 to 15 basis points.

3.3 Pricing of sig-payoffs

The tractability of the model is also crucial when it comes to pricing, in particular of path-dependent options. Indeed, sig-payoffs, like Asian forwards, can be priced via the following formula. For generic payoffs these sig-payoffs can be used in an approximate manner, e.g. similarly as in Ackerer and Filipović (2020) for standard polynomial processes, or as control variates to reduce the variance in Monte-Carlo pricing.

Theorem 1. The price of a sig-payoff $\langle e_J, \widehat{\mathbb{S}}_{\mathbb{T}}(\ell) \rangle$ can be expressed as

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[\langle e_J, \widehat{\mathbb{S}}_T(\ell) \rangle] &= \langle e(J,\ell), \mathbb{E}_{\mathbb{Q}}[\widehat{\mathbb{X}}_T] \rangle \\ &= \sum_I p_I(J,\ell) \langle e_I, \mathbb{E}_{\mathbb{Q}}[\widehat{\mathbb{X}}_T] \rangle, \end{aligned}$$

where $p_I(J,\ell)$ are polynomials in the coefficients of ℓ.

As already stated above to compute $\mathbb{E}_{\mathbb{Q}}[\widehat{\mathbb{X}}_T]$, an affine and polynomial process point view works for generic primary processes $\widehat{X}$ of form (1) (see Cuchiero et al. (2022) for details).

REFERENCES

Ackerer, D. and Filipović, D. (2020). Option pricing with orthogonal polynomial expansions. *Mathematical Finance*, 30(1), 47–84.

Akyildirim, E., Gambara, M., Teichmann, J., and Zhou, S. (2022). Applications of Signature Methods to Market Anomaly Detection. *preprint arXiv:2201.02441*.

Bayer, C., Hager, P., Riedel, S., and Schoenmakers, J. (2021). Optimal stopping with signatures. *preprint arXiv:2105.00778*.

Buehler, H., Horvath, B., Lyons, T., Perez Arribas, I., and Wood, B. (2020). A data-driven market simulator for small data environments. *Available at SSRN 3632431*.

Chen, K. (1957). Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Annals of Mathematics*, 163–178.

Chen, K. (1977). Iterated path integrals. *Bulletin of the American Mathematical Society*, 831–879.

Cohen, S.N., Reisinger, C., and Wang, S. (2021). Arbitrage-free neural-SDE market models. *preprint arXiv:2105.11053*.

Cuchiero, C., Khosrawi, W., and Teichmann, J. (2020). A generative adversarial network approach to calibration of local stochastic volatility models. *Risks*, 8(4), 101.

Cuchiero, C., Keller-Ressel, M., and Teichmann, J. (2012). Polynomial processes and their applications to mathematical finance. *Finance and Stochastics*, 16(4), 711–740.

Cuchiero, C., Svaluto-Ferro, S., and Teichmann, J. (2022). Signature SDEs from an affine and polynomial perspective. *In preparation*.

Duffie, D., Filipović, D., and Schachermayer, W. (2003). Affine processes and applications in finance. *Annals of Applied Probability*, 13, 984–1053.

Gierjatowicz, P., Sabate-Vidales, M., Siska, D., Szpruch, L., and Zuric, Z. (2020). Robust pricing and hedging via neural SDEs. *Available at SSRN 3646241*.

Lyons, T.J. (1998). Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2), 215–310.

Lyons, T., Nejad, S., and Perez Arribas, I. (2020). Nonparametric pricing and hedging of exotic derivatives. *Applied Mathematical Finance*, 27(6), 457–494.

Ni, H., Szpruch, L., Sabate-Vidales, M., Xiao, B., Wiese, M., and Liao, S. (2021). Sig-Wasserstein GANs for Time Series Generation. *preprint arXiv:2111.01207*.

Perez Arribas, I. (2020). *Signatures in machine learning and finance*. Ph.D. thesis, University of Oxford.

Perez-Arribas, I., Salvi, C., and Szpruch, L. (2020). Sig-SDEs model for quantitative finance.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Code-based digital signatures: state of the art and open challenges

Marco Baldi *

* *Università Politecnica delle Marche, Ancona, Italy (e-mail: m.baldi@univpm.it).*

Abstract: The problem of decoding a random-like linear block code is recognized as one of the most important mathematical problems that apparently will remain hard even with the availability of solvers based on quantum computers. This motivates an increasing interest in code-based cryptography as a solution for the design of post-quantum cryptographic primitives. However, while several robust and efficient code-based systems exist for asymmetric encryption and key exchange, mostly stemming from the McEliece and Niederreiter original cryptosystems, devising robust and efficient code-based signature schemes is a far more challenging task. This work provides an overview of past and current approaches to the problem of designing secure and practical code-based signature schemes following two main directions: adapting the McEliece and Niederreiter schemes to the digital signature setting following the classical hash-and-sign approach or deriving digital signatures from code-based identification schemes through suitable transformations.

Keywords: Code-based cryptography, decoding problem, digital signatures, identification schemes, post-quantum cryptography.

1. INTRODUCTION

Digital signatures traditionally rely on cryptographic trapdoors, that is, functions whose solution is easy to compute for those who own some secret, known as private key, while requiring the solution of some computationally hard problem for all the others. A public key is then derived from the private key through some sort of one-way function, such that it can be publicly distributed without revealing the private key. In the case of digital signatures, the private key of a user allows computing their digital signature of a message, while the associated public key allows verifying the signature or, equivalently, ascertaining the signer's identity and the message integrity.

Cryptographic trapdoors used in widespread digital signature schemes rely on the hardness of either factoring large semiprime numbers or computing discrete logarithms. These classical hard problems have withstood cryptanalysis for about fifty years, and have enabled very fast digital signature schemes with compact signatures. However, quantum algorithms like Shor's algorithm (Shor, 1997) make the same problems solvable in polynomial time with a quantum computer, which motivates the search for alternative, quantum-resistant solutions (Moody, 2021).

The most established solutions for constructing quantum-resistant cryptographic trapdoors are those relying on lattice-based problems, code-based problems, multivariate polynomial problems, isogenies between elliptic curves and others (Chen et al., 2016). Differently from problems based on lattices, which have already been exploited for constructing both efficient asymmetric encryption and digital signature schemes, problems based on coding are more challenging to be exploited as a basis for digital signature schemes. This work provides a critical overview of current solutions for post-quantum digital signatures based on codes, focusing on challenging aspects and describing the most promising avenues for the development of efficient code-based digital signature schemes.

2. HARDNESS OF THE DECODING PROBLEM

The main hard problem derived from coding theory is the so-called decoding problem, that is, the problem of finding the nearest vector (or codeword) belonging to a k-dimensional subspace (named code) $\mathcal{C} \subset \mathbb{F}_q^n$ starting from any vector $\mathbf{x} \in \mathbb{F}_q^n$. Any $k \times n$ matrix $\mathbf{G}$ over $\mathbb{F}_q$ forming a basis for $\mathcal{C}$ is called generator matrix for the code $\mathcal{C}$, while any $(n-k) \times n$ matrix $\mathbf{H}$ over $\mathbb{F}_q$ forming a basis for its dual $\mathcal{C}^\perp$ is called a parity-check matrix for $\mathcal{C}$, such that $\mathcal{C}$ is the kernel of $\mathbf{H}$. The parity-check matrix $\mathbf{H}$ divides the space $\mathbb{F}_q^n$ into q^{n-k} cosets, that is, any vector $\mathbf{x} \in \mathbb{F}_q^n$ belongs to the coset represented by $\mathbf{H} \cdot \mathbf{x}^T$, also known as the syndrome of $\mathbf{x}$ through $\mathbf{H}$. Since any codeword belongs to the coset corresponding to the all-zero syndrome, the aforementioned decoding problem can also be formulated in terms of syndrome, that is, finding the smallest weight vector $\mathbf{e}$ belonging to the same coset of $\mathbf{x}$. In such a case, due to the code linearity, $\mathbf{x} - \mathbf{e}$ has an all-zero syndrome and thus belongs to the code, actually being the decoded codeword. This alternative formulation of the decoding problem is known as syndrome decoding problem.

If we consider the Hamming metric, that is, we define syndrome decoding as finding the vector with the smallest Hamming weight in a coset, then the problem has been proved to be NP-complete for random codes (Berlekamp et al., 1978; Barg, 1994). A number of algorithms have been developed over years for solving such a problem

(Prange, 1962; Dumer, 1989; Lee and Brickell, 1988; Leon, 1988; Stern, 1988; Becker et al., 2012; May et al., 2011; Bernstein et al., 2011), all characterized by exponential complexity, although with progressively decreasing exponential factors.

3. DIGITAL SIGNATURES BASED ON HASH-AND-SIGN

The first and traditional approach to code-based cryptography is the one introduced by McEliece in 1978 (McEliece, 1978), according to which the cleartext message is encoded into a codeword of a public code, and then corrupted with a number of intentional errors below the error correction capability of the public code. The trapdoor is constructed by providing the receiver with a secret representation of the public code which enables decoding through efficient algorithms, while the corresponding public representation of the same codes forces the use of algorithms for general decoding of random codes, which are characterized by exponential complexity. An alternative formulation of the same approach has been introduced by Niederreiter in 1986 (Niederreiter, 1986), and exploits syndromes in the place of noisy codewords.

While these approaches based on decoding work very well for asymmetric encryption, leading to robust and efficient public-key cryptosystems, their conversion into digital signature schemes is more challenging. In fact, differently from some asymmetric cryptosystems like RSA, these cryptosystems cannot be easily converted into hash-and-sign digital signature schemes. In fact, opposed to RSA, these cyptosystems have a domain and an image of the encryption function that do not coincide, which makes them difficult to be "reversed" for obtaining a hash-and-sign scheme. In fact, it is very unlikely that a random cleartext message, or its hash digest, coincides with a syndrome that is correctable through the secret code, or with a codeword of the same code affected by a limited number of errors.

Thus, some workaround needs to be used for making signatures of random messages or their digests easy to generate in the traditional code-based setting, and this normally undermines the primitive security. A well-known proposal in this sense has been made by Courtois, Finiasz, and Sendrier (Courtois et al., 2001), but it has been found exposed to some vulnerabilities that require the adoption of very costly solutions from the point of view of signature generation time and size of public keys. The use of codes characterized by sparse representations, which is a promising avenue for achieving public-key encryption schemes derived from McEliece with compact keys, has also been attempted to build signature schemes relying on the hardness of decoding (Baldi et al., 2013), but with little fortune (Phesso and Tillich, 2016).

A more recent approach is that of exploiting large-weight error vectors, instead of small-weight ones, which are also difficult to find through decoding, especially over non-binary fields. The scheme proposed in (Debris-Alazard et al., 2019), named Wave, exploits this fact, and uses codes in the $(U|U+V)$ form to enable aided versions of general decoding algorithms for signature generation. This provides a step further with respect to CFS, by achieving a public key size growing as λ^2, where λ is the security level in bits. However, a public key with size in the order of 4 megabytes is required for achieving 128 bits of classical security, which is rather large. Moreover, generating signatures requires the execution of a modified version of Prange's general decoding algorithm (Prange, 1962), with a complexity in the order of λ^3, thus resulting in a slow signature generation. Another scheme relying on the hardness of decoding of large-weight vectors has been recently introduced in (Baldi et al., 2022), with the main advantage of a fast signature generation, thanks to the possibility of generating signatures without requiring the execution of a decoding algorithm. However, secure instances have still been studied only for one-time use, and further investigations are needed to address the multiple-time use case.

All the above schemes rely on some hidden code structure, which certainly represents a delicate point and can turn into a vulnerability if the structure of the secret code is not adequately hidden. Such a hidden code structure is no longer required in code-based signature schemes derived from identification schemes, as described next.

4. DIGITAL SIGNATURES FROM IDENTIFICATION SCHEMES

An alternative approach to design digital signature schemes is the one stemming from code-based identification schemes. Classical identification protocols based on zero-knowledge proofs allow one party (the prover) to demonstrate to another party (the verifier) knowledge about a secret, without revealing the secret itself, and work as follows.

(1) The prover, owning the secret, commits to some random data that is sent to the verifier.
(2) The verifier receives the random data and generates a corresponding challenge for the prover.
(3) The prover generates a response to the challenge, without revealing anything concerning the secret.

A zero knowledge identification protocol of this type can be converted into a digital signature scheme through the Fiat-Shamir approach (Fiat and Shamir, 1986), according to which the challenge is obtained through a deterministic function applied to the message, besides committed data, and the transcript of the protocol provides the signature. The rationale of the Fiat-Shamir approach is that of replacing the prover's choices with the output of a hash-based algorithm, thus making the protocol no longer interactive and suitable for digital signatures. The verifier, owning the public key, can then use it to verify that the transcript is actually consistent with the message, thus validating the signature.

The first code-based identification scheme has been proposed by Stern in 1994 (Stern, 1994); several variants have subsequently appeared in the following years (e.g., (Véron, 1997; Gaborit and Girault, 2007; Cayrel et al., 2011; El Yousfi Alaoui et al., 2013)). The main advantage of these schemes is that they use public random-like codes without any hidden structure, which eliminates an important source of potential vulnerabilities. Hence, security is based on a pure, random instance of some (usually NP-hard) problem. These schemes are also characterized by

very compact public keys, which is another important advantage.

However, for many years, digital signature schemes derived from identification protocols have been considered not practical, because of their normally large signatures. In fact, these schemes are characterized by a usually high soundness error, which implies that an adversary can cheat on a single execution of the protocol with non-trivial probability, for example 1/2 or 2/3. This requires iterating the protocol for a significant number of times in order for the adversary to experience a sufficiently small overall cheating probability (say, not greater than $2^{-\lambda}$). As a consequence, the transcript of the protocol becomes rather long, which translates into a significant size of generated signatures, in the order of tens or even hundreds of kilobytes.

However, some recent works have described how to apply several optimizations on top of an identification scheme, in order to reduce the signature size (Gueron et al., 2022; Bettaieb et al., 2021; Bidoux et al., 2022; Becker et al., 2020; Barenghi et al., 2021; Feneuil et al., 2021). These techniques can be thought of as a clever rewriting of the straight identification protocol, with no impact on the underlying security assumptions. Some of them consist, for instance, in extracting randomness from a so-called seed tree, or emulating an imaginary multiparty computation phase. Techniques of this kind have been shown to have a great potential in reducing the signature size. For this reason, zero knowledge identification schemes are currently deemed as one of the most promising strategies to achieve secure and efficient code-based signatures. Another promising avenue to achieve compact public keys and reduced communication costs in zero-knowledge identification schemes based on codes is that of restricting the entries of the searched vector within a subset of the underlying finite field (Baldi et al., 2020).

In this context, but under the setting of lattice-based problems, an important advance is represented by the Schnorr-Lyubashevsky approach (Lyubashevsky, 2012), which indeed allows deriving compact digital signatures from identification schemes. Some attempts have been made for translating such an approach in the code-based setting (Persichetti, 2018), but the resulting schemes have been successfully exposed to cryptanalysis based on statistical approaches (Aragon et al., 2021).

4.1 Identification schemes based on code equivalence

In 2020, a new code-based zero knowledge identification scheme named LESS has been proposed (Biasse et al., 2020). Differently from classical code-based identification schemes, it comes as a three pass protocol in which the soundness error can be arbitrarily reduced by using multiple key pairs. These characteristics are intrinsically due to the fact that the scheme is constructed upon the somehow unorthodox code equivalence problem (which can be described as a transitive code-based group action). In a nutshell, the problem reads as follows: given two linear codes, find an isometry which maps one code into the other. Notably, determining whether two codes are equivalent is a standard problem in coding theory and, as such, has been studied for decades. Despite the fact that code-equivalence is not NP-hard (unless the polynomial hierarchy collapses), there exist instances (e.g., monomial equivalences or permutations of self-dual codes) for which no efficient solver is known. Interestingly, for such instances, the currently known best attack (Beullens, 2020) reduces to the problem of decoding a linear code, which is instead NP-complete, as already said. The original LESS parameters have been broken in (Beullens, 2020), but in a subsequent paper new and secure instances have been recommended (Barenghi et al., 2021), together with several optimizations to make signatures more compact.

REFERENCES

Aragon, N., Baldi, M., Deneuville, J.C., Khathuria, K., Persichetti, E., and Santini, P. (2021). Cryptanalysis of a code-based full-time signature. *Designs, Codes and Cryptography*, 89(9), 2097–2112.

Baldi, M., Battaglioni, M., Chiaraluce, F., Horlemann-Trautmann, A., Persichetti, E., Santini, P., and Weger, V. (2020). A new path to code-based signatures via identification schemes with restricted errors. *CoRR*, abs/2008.06403. URL `https://arxiv.org/abs/2008.06403`.

Baldi, M., Bianchi, M., Chiaraluce, F., Rosenthal, J., and Schipani, D. (2013). Using LDGM codes and sparse syndromes to achieve digital signatures. In P. Gaborit (ed.), *Post-Quantum Cryptography*, 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg.

Baldi, M., Chiaraluce, F., and Santini, P. (2022). SPANSE: combining sparsity with density for efficient one-time code-based digital signatures. *CoRR*, abs/2205.12887. URL `https://arxiv.org/abs/2205.12887`.

Barenghi, A., Biasse, J.F., Persichetti, E., and Santini, P. (2021). LESS-FM: fine-tuning signatures from the code equivalence problem. In J.H. Cheon and J.P. Tillich (eds.), *Post-Quantum Cryptography - 12th International Workshop, PQCrypto 2021*, volume 12841 of *Lecture Notes in Computer Science*, 23–43. Springer.

Barg, S. (1994). Some new NP-complete coding problems. *Problemy Peredachi Informatsii*, 30(3), 23–28.

Becker, A., Joux, A., May, A., and Meurer, A. (2012). Decoding random binary linear codes in $2^{n/20}$: How 1+1=0 improves information set decoding. In D. Pointcheval and T. Johansson (eds.), *Advances in Cryptology - EUROCRYPT 2012*, volume 7237 of *LNCS*, 520–536. Springer.

Becker, A., Joux, A., May, A., and Meurer, A. (2020). Sigma protocols for MQ, PKP and SIS, and fishy signature schemes. In A. Canteaut and Y. Ishai (eds.), *Advances in Cryptology – EUROCRYPT 2020*, volume 12107 of *Lecture Notes in Computer Science*, 183–211. Springer, Cham.

Berlekamp, E., McEliece, R., and van Tilborg, H. (1978). On the inherent intractability of certain coding problems. *IEEE Trans. Inf. Theory*, 24(3), 384–386.

Bernstein, D.J., Lange, T., and Peters, C. (2011). Smaller decoding exponents: ball-collision decoding. In *Annual Cryptology Conference*, 743–760. Springer.

Bettaieb, S., Bidoux, L., Blazy, O., and Gaborit, P. (2021). Zero-knowledge reparation of the Véron and AGS code-based identification schemes. In *Proc. 2021 IEEE International Symposium on Information Theory (ISIT 2021)*, 55–60. Melbourne, Victoria, Australia.

Beullens, W. (2020). Not Enough LESS: An Improved Algorithm for Solving Code Equivalence Problems over $\mathbb{F}_q$. In *International Conference on Selected Areas in Cryptography*, 387–403. Springer.

Biasse, J.F., Micheli, G., Persichetti, E., and Santini, P. (2020). LESS is More: Code-Based Signatures Without Syndromes. In A. Nitaj and A. Youssef (eds.), *Progress in Cryptology - AFRICACRYPT 2020*, 45–65. Springer International Publishing, Cham.

Bidoux, L., Gaborit, P., Kulkarni, M., and Mateu, V. (2022). Code-based signatures from new proofs of knowledge for the syndrome decoding problem. *CoRR*, abs/2201.05403. URL `https://arxiv.org/abs/2201.05403`.

Cayrel, P.L., Véron, P., and El Yousfi Alaoui, S.M. (2011). A zero-knowledge identification scheme based on the q-ary syndrome decoding problem. In *International Conference on Selected Areas in Cryptography*, 171–186. Springer Berlin Heidelberg.

Chen, L., Jordan, S., Liu, Y.K., Moody, D., Peralta, R., Perlner, R., and Smith-Tone, D. (2016). Report on post-quantum cryptography. Technical Report NISTIR 8105, National Institute of Standards and Technology.

Courtois, N., Finiasz, M., and Sendrier, N. (2001). How to achieve a McEliece-based digital signature scheme. In *ASIACRYPT*, 157–174.

Debris-Alazard, T., Sendrier, N., and Tillich, J.P. (2019). Wave: A new family of trapdoor one-way preimage sampleable functions based on codes. In *ASIACRYPT*, 21–51. Springer.

Dumer, I.I. (1989). Two decoding algorithms for linear codes. *Problemy Peredachi Informatsii*, 25(1), 24–32.

El Yousfi Alaoui, S.M., Cayrel, P.L., El Bansarkhani, R., and Hoffmann, G. (2013). Code-based identification and signature schemes in software. In A. Cuzzocrea, C. Kittl, D.E. Simos, E. Weippl, and L. Xu (eds.), *Security Engineering and Intelligence Informatics*, 122–136. Springer Berlin Heidelberg, Berlin, Heidelberg.

Feneuil, T., Joux, A., and Rivain, M. (2021). Shared permutation for syndrome decoding: New zero-knowledge protocol and code-based signature. Cryptology ePrint Archive, Report 2021/1576. `https://ia.cr/2021/1576`.

Fiat, A. and Shamir, A. (1986). How to prove yourself: Practical solutions to identification and signature problems. In A.M. Odlyzko (ed.), *Advances in Cryptology — CRYPTO' 86 Proceedings. CRYPTO 1986*, volume 263 of *Lecture Notes in Computer Science*, 186–194. Springer, Berlin, Heidelberg.

Gaborit, P. and Girault, M. (2007). Lightweight code-based identification and signature. In *2007 IEEE International Symposium on Information Theory*, 191–195. IEEE.

Gueron, S., Persichetti, E., and Santini, P. (2022). Designing a practical code-based signature scheme from zero-knowledge proofs with trusted setup. *Cryptography*, 6(1:5).

Lee, P.J. and Brickell, E.F. (1988). An observation on the security of McEliece's public-key cryptosystem. In *Workshop on the Theory and Application of Cryptographic Techniques*, 275–280. Springer.

Leon, J.S. (1988). A probabilistic algorithm for computing minimum weights of large error-correcting codes. *IEEE Trans. Inf. Theory*, 34(5), 1354–1359.

Lyubashevsky, V. (2012). Lattice signatures without trapdoors. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 738–755. Springer.

May, A., Meurer, A., and Thomae, E. (2011). Decoding random linear codes in $\mathcal{O}(2^{0.054n})$. In *International Conference on the Theory and Application of Cryptology and Information Security*, 107–124. Springer.

McEliece, R.J. (1978). A public-key cryptosystem based on algebraic coding theory. *DSN Progress Report*, 114–116.

Moody, D. (2021). Status update on the 3rd round. Technical report, NIST. URL `https://csrc.nist.gov/presentations/2021/status-update-on-the-3rd-round`.

Niederreiter, H. (1986). Knapsack type cryptosystems and algebraic coding theory. *Problems of Control and Information Theory. Problemy Upravlenija i Teorii Informacii*, 15, 19–34.

Persichetti, E. (2018). Efficient one-time signatures from quasi-cyclic codes: A full treatment. *Cryptography*, 2(4:30). doi:10.3390/cryptography2040030.

Phesso, A. and Tillich, J.P. (2016). An efficient attack on a code-based signature scheme. In T. Takagi (ed.), *Post-Quantum Cryptography*, 86–103. Springer International Publishing, Cham.

Prange, E. (1962). The use of information sets in decoding cyclic codes. *IRE Trans. Inf. Theory*, 8(5), 5–9.

Shor, P. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5), 1484–1509.

Stern, J. (1988). A method for finding codewords of small weight. In *International Colloquium on Coding Theory and Applications*, 106–113. Springer.

Stern, J. (1994). A new identification scheme based on syndrome decoding. In D.R. Stinson (ed.), *Advances in Cryptology — CRYPTO' 93*, 13–21. Springer Berlin Heidelberg.

Véron, P. (1997). Improved identification schemes based on error-correcting codes. *Applicable Algebra in Engineering, Communication and Computing*, 8(1), 57–69.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On representations of port-Hamiltonian DAE systems

Volker Mehrmann * **Arjan van der Schaft** **

* *Institut für Mathematik, TU-Berlin*
(e-mail: mehrmann@math.tu-berlin.de)
** *Bernoulli Institute for Mathematics, Computer Science and AI,*
Jan C. Willems Center for Systems and Control,
University of Groningen (e-mail: a.j.van.der.schaft@rug.nl)

Abstract: Port-Hamiltonian DAE systems are discussed that are the composition of a Dirac structure and a Lagrangian subspace, where the latter is generalizing the Hamiltonian function expressing energy storage. The algebraic constraints in such systems are correspondingly divided into Dirac and Lagrange algebraic constraints. The relations between different representations of the same port-Hamiltonian DAE system are discussed.

Keywords: Port-Hamiltonian systems, algebraic constraints, Dirac structures, Lagrangian subspaces, equivalence

1. PORT-HAMILTONIAN DAE SYSTEMS

Differential-algebraic equation (DAE) system models are ubiquitous in physical systems modeling. In particular, the algebraic constraints often arise from interconnection of subsystems. Port-based modeling of complex engineering systems leads to a broad, but specific, class of DAE systems, called *port-Hamiltonian DAE systems*. Analysis, simulation, and control of such systems can benefit from a closer study of the structural properties of port-Hamiltonian DAE systems.

From a port-based modeling perspective Van der Schaft (2013, 2017); Van der Schaft, Jeltsema (2014), the algebraic constraints in port-Hamiltonian systems are primarily reflected in the Dirac structure of the system. However, as was argued before, e.g. in Beattie et al. (2017, 2019); Mehrmann et al. (2018), and more explicitly in Barbero-Linan et al. (2019); Van der Schaft, Maschke (2018, 2020); Gernandt et al. (2021), algebraic constraints may also arise by replacing the Hamiltonian function in the definition of port-Hamiltonian systems by a *Lagrangian subspace* (more general than the graph of the gradient of the Hamiltonian function). The present contribution aims at taking a closer look at the representations and properties of such generalized port-Hamiltonian DAE systems. For simplicity of exposition we will concentrate on linear port-Hamiltonian DAE systems *without* energy dissipation and *without* inputs and outputs. The remaining (autonomous) port-Hamiltonian DAE systems are fully described by two geometric structures, namely a Lagrangian subspace and a Dirac structure.

Recall the definitions of such structures. A Dirac structure on an n-dimensional linear state space $\mathcal{X}$ is specified by a subspace $\mathcal{D} \subset \mathcal{X} \times \mathcal{X}^*$, which is a maximal subspace on which the symmetric canonical bilinear form on $\mathcal{X} \times \mathcal{X}^*$ represented by the $2n \times 2n$ matrix

$$\begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \tag{1}$$

is zero. It follows that the dimension of any Dirac structure $\mathcal{D}$ is equal to $\dim \mathcal{X} = n$, and that there exist $n \times n$ matrices K and L such that

$$\mathcal{D} = \ker \begin{bmatrix} K & L \end{bmatrix} = \operatorname{im} \begin{bmatrix} L^\top \\ K^\top \end{bmatrix} \subset \mathcal{X} \times \mathcal{X}^*, \quad KL^\top + LK^\top = 0, \tag{2}$$

while conversely any such pair (K, L) defines a Dirac structure $\mathcal{D}$. Note that the property $KL^\top + LK^\top = 0$ is a generalized skew-symmetry property, and reflects the fact that $e^\top f = 0$ for any $(f, e) \in \mathcal{D}$, expressing power conservation.

Analogously, a Lagrangian subspace is a maximal subspace on which the skew-symmetric canonical symplectic form on $\mathcal{X} \times \mathcal{X}^*$ represented by the $2n \times 2n$ matrix

$$\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \tag{3}$$

is zero. The dimension of any Lagrangian subspace $\mathcal{L}$ is equal to $\dim \mathcal{X} = n$, and for any $\mathcal{L}$ there exist $n \times n$ matrices P and S such that

$$\mathcal{L} = \ker \begin{bmatrix} -S^\top & P^\top \end{bmatrix} = \operatorname{im} \begin{bmatrix} P \\ S \end{bmatrix} \subset \mathcal{X} \times \mathcal{X}^*, \quad S^\top P = P^\top S, \tag{4}$$

while conversely any such pair (P, S) defines a Lagrangian subspace $\mathcal{L}$. The Lagrangian subspace corresponding to a quadratic Hamiltonian function $H(x) = \frac{1}{2} x^\top Q x$, $Q = Q^\top$, is defined by $P = I, S = Q$.

The port-Hamiltonian DAE system corresponding to a pair $(\mathcal{D}, \mathcal{L})$ of a Dirac structure and Lagrangian subspace is defined by the *composition* of $\mathcal{D}$ and $\mathcal{L}$ (over the shared variables $e \in \mathcal{X}^*$) , that is

$$\begin{aligned} \mathcal{D} \circ \mathcal{L} = \{(f, x) \in \mathcal{X} \times \mathcal{X} \mid \text{there exists } e \in \mathcal{X}^* \text{ such that} \\ (f, e) \in \mathcal{D}, (e, x) \in \mathcal{L}\} \end{aligned} \tag{5}$$

This defines the port-Hamiltonian dynamics

$$(-\dot{x}, e) \in \mathcal{D} \circ \mathcal{L} \tag{6}$$

(Note that strictly speaking $-\dot{x} = f$ is an element of the *tangent space* of $\mathcal{X}$ at $x \in \mathcal{X}$, which however can be identified with $\mathcal{X}$.) Coordinate representations of the port-Hamiltonian dynamics (6) can be obtained as follows.

1.1 The x-representation

The composition $\mathcal{D} \circ \mathcal{L} \subset \mathcal{X} \times \mathcal{X}$ can be explicitly computed as follows. First consider the 'intersection' of $\mathcal{D}$ and $\mathcal{L}$, given by the kernel of

$$\begin{bmatrix} K & L & 0 \\ 0 & -P^\top & S^\top \end{bmatrix} \tag{7}$$

Consider a maximal annihilator $[M \;\; N]$ of $\begin{bmatrix} L \\ -P^\top \end{bmatrix}$, that is

$$\ker [M \;\; N] = \operatorname{im} \begin{bmatrix} L \\ -P^\top \end{bmatrix} \tag{8}$$

and thus in particular $ML - NP^\top = 0$. Premultiply the 'intersection' with this maximal annihilator, so as to obtain

$$[M \;\; N] \begin{bmatrix} K & L & 0 \\ 0 & -P^\top & S^\top \end{bmatrix} = \begin{bmatrix} MK & 0 & NS^\top \end{bmatrix} \tag{9}$$

Then $\mathcal{D} \circ \mathcal{L} \subset \mathcal{X} \times \mathcal{X}$ is given as

$$\mathcal{D} \circ \mathcal{L} = \ker \begin{bmatrix} MK & NS^\top . \end{bmatrix} \tag{10}$$

Thus the resulting DAE system (in the original state variables x for $\mathcal{X}$) is the port-Hamiltonian DAE system

$$MK\dot{x} = NS^\top x, \quad \text{where } ML = NP^\top. \tag{11}$$

Remark 1.1. If P is invertible, then also M is invertible (and conversely). In this case we may take $M = I$ and N such that

$$\ker [I \;\; N] = \operatorname{im} \begin{bmatrix} L \\ -P^\top \end{bmatrix}, \tag{12}$$

implying $L = NP^\top$. Thus $N = LP^{-\top}$ and the DAE system takes the form

$$K\dot{x} = LP^{-\top}S^\top x = L\left(SP^{-1}\right)^\top x = LSP^{-1}x, \tag{13}$$

where the last equality follows from $S^\top P = P^\top S$. This is exactly the form of a port-Hamiltonian DAE system in case of a general Dirac subspace, and a Lagrangian subspace which is given as the graph of the symmetric matrix $Q = SP^{-1}$. If additionally K is invertible then we obtain the Poisson formulation of Hamiltonian dynamics

$$\dot{x} = \left(K^{-1}L\right)\left(SP^{-1}\right)x, \tag{14}$$

where the skew-symmetric matrix $J := K^{-1}L$ defines a Poisson structure.

Remark 1.2. If instead we assume L to be invertible, then N is invertible, and thus we can take $N = I$. In this case the equations become

$$MK\dot{x} = S^\top x, \quad ML = P^\top.$$

Substituting $M = P^\top L^{-1}$ we obtain the port-Hamiltonian DAE

$$P^\top L^{-1} K\dot{x} = S^\top x \tag{15}$$

where $\left(L^{-1}K\right)^\top = -L^{-1}K$. If additionally P is invertible, this may be rewritten as

$$\left(L^{-1}K\right)\dot{x} = \left(SP^{-1}\right)x, \tag{16}$$

which is the standard symplectic formulation of Hamiltonian dynamics in case the skew-symmetric matrix $L^{-1}K$ is invertible.

1.2 The z-representation

Another way to obtain a DAE representation of the defining pair $(\mathcal{D}, \mathcal{L})$ is to consider a *parametrization* of the Lagrangian subspace

$$\begin{bmatrix} x \\ e \end{bmatrix} = \begin{bmatrix} P \\ S \end{bmatrix} z \tag{17}$$

with $z \in \mathcal{Z}$, where $\mathcal{Z}$ is an n-dimensional parametrization space. As detailed in Van der Schaft, Maschke (2018) this yields the port-Hamiltonian DAE system

$$KP\dot{z} = LSz \tag{18}$$

Note that singularity of the matrix KP (and thus the appearance of algebraic constraints) may originate from two different sources: singularity of K (corresponding to *Dirac algebraic constraints*), as well as singularity of P (*Lagrange algebraic constraints*).

The Hamiltonian of the port-Hamiltonian DAE system (18) is given as

$$\mathcal{H}^z(z) = \frac{1}{2} z^\top S^\top P z, \tag{19}$$

and indeed $\frac{d}{dt}\mathcal{H}^z(z) = z^\top S^\top P\dot{z} = 0$ by the property $e^\top f = 0$ for any $(f, e) \in \mathcal{D}$. In case P is invertible the Hamiltonian $\mathcal{H}^x$ of the x-representation is

$$\mathcal{H}^x(x) = \frac{1}{2} x^\top SP^{-1}x \tag{20}$$

Substituting $x = Pz$ we immediately observe the following equality with $\mathcal{H}^z$:

$$\mathcal{H}^x(x) = \frac{1}{2} x^\top SP^{-1}x = \frac{1}{2} z^\top P^\top SP^{-1}Pz = \mathcal{H}^z(z). \tag{21}$$

Alternatively, starting from the x-representation we can define in case S is invertible the *co-energy* (Legendre transform)

$$\mathcal{H}_c^x(e) = \frac{1}{2} e^\top PS^{-1}e, \tag{22}$$

for which $\mathcal{H}_c^x(e) = \mathcal{H}^z(z)$ for $e = Sz$.

Note that the state vector x of the x-representation (11) is in general (especially if P is singular) *different* from the state vector of the z-representation (18); although of equal dimension. In the presentation and in the forthcoming paper Mehrmann, van der Schaft (2022) the precise notion of equivalence of these two representations will be discussed. Furthermore, we aim to address the structural properties of port-Hamiltonian DAE systems, including index analysis.

2. WHEN IS A DAE SYSTEM A PORT-HAMILTONIAN DAE SYSTEM: A GENERALIZED LYAPUNOV EQUATION

Next we study the question when a general DAE system $\Sigma \subset \mathcal{X} \times X$ given as $E\dot{x} = Ax$, $x \in \mathcal{X}$, is actually a port-Hamiltonian system; i.e., when does there exist a Dirac structure $\mathcal{D} \subset \mathcal{X} \times \mathcal{X}^*$ and a Lagrangian subspace $\mathcal{L} \subset \mathcal{X} \times \mathcal{X}^*$ such that $\Sigma = \mathcal{D} \circ \mathcal{L}$. First we identify two *necessary* conditions for the existence of $(\mathcal{D}, \mathcal{L})$.

Proposition 2.1. Consider $\Sigma \subset \mathcal{X} \times \mathcal{X}$ represented by $Ef + Ax = 0$ and the Lagrangian subspace $\mathcal{L} \subset \mathcal{X} \times \mathcal{X}^*$ represented by P, S. Then necessary conditions for the existence of a subspace $\mathcal{D}$ such that $\Sigma = \mathcal{D} \circ \mathcal{L}$ are

$$\begin{aligned} &(i)\ \ker S^\top \subset \ker A \\ &(ii) A^{-1}(\operatorname{im} E) \subset \operatorname{im} P \end{aligned} \tag{23}$$

Next step will be to develop a *Lyapunov equation* (in terms of the unknown pair (P, S) defining a Lagrange subspace). This is aimed at further generalizing the theory of Lyapunov equations for DAE systems developed in e.g. Stykel (2002); Reis et al. (2015).

REFERENCES

M. Barbero-Linan, H. Cendra, E. Garcia-Torano Andres, D. Martin de Diego, Morse families and Dirac systems, *Journal of Geometric Mechanics*, 11(4), 487–510, 2019.

C.A. Beattie, V. Mehrmann, H. Xu, H. Zwart, Port-Hamiltonian descriptor systems, *Mathematics of Control Signals and Systems*, 30(4), 2017.

C. A. Beattie, V. Mehrmann, P. Van Dooren, Robust port-Hamiltonian representations of passive systems, *Automatica*, 100, 182–186, 2019.

H. Gernandt, F.E. Haller, T. Reis, A linear relation approach to port-Hamiltonian differential-algebraic equations, *SIAM Journal on Matrix Analysis and Applications*, 42(2),1011–1044, 2021.

V. Mehrmann, C. Mehl, M. Wojtylak, Linear algebra properties of dissipative Hamiltonian descriptor systems, *SIAM Journal on Matrix Analysis and Applications*, 39(3), 1489–1519, 2018.

V. Mehrmann, A.J. van der Schaft, in preparation, 2022.

T. Reis, O. Rendel, M. Voigt, The KalmanYakubovich-Popov inequality for differential-algebraic systems, *Linear Algebra and its Applications* 485, 153-193, 2015.

T. Stykel, Stability and inertia theorems for generalized Lyapunov equations, *Linear Algebra and its Applications* 355 (1-3), 297–314, 2002.

A. J. van der Schaft, Port-Hamiltonian differential-algebraic systems. In *Surveys in Differential-Algebraic Equations I*, pp. 173 – 226. Springer-Verlag, 2013.

A.J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, 3rd Edition 2017, Springer International.

A.J. van der Schaft, D. Jeltsema, "Port-Hamiltonian Systems Theory: An Introductory Overview," *Foundations and Trends in Systems and Control*, 1(2/3), 173–378, 2014.

A.J. van der Schaft, B. Maschke, Generalized port-Hamiltonian DAE systems, *Systems & Control Letters*, 121, 31–37, 2018.

A.J. van der Schaft, B. Maschke, Dirac and Lagrange algebraic constraints in nonlinear port-Hamiltonian systems, *Vietnam Journal of Mathematics*, 48(4), 929–939, 2020.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Computational Methods for Uniform Ensemble Reachability ⋆

Michael Schönlein *

* *Faculty of Computer Science and Mathematics, University of Passau, Germany*

Abstract: We consider reachability properties of families of parameter-dependent linear systems, where the inputs are restricted to be independent of the parameter. If for every family of parameter-dependent target states and every neighborhood of it there is an input such that the zero state can be steered simultaneously into the given neighborhood the parameter-dependent system is called ensemble reachable. Recently, a lot of effort has been spent on the derivation of necessary and sufficient conditions for ensemble reachability. Here we tackle the subsequent question how to determine a suitable input if the target family and the neighborhood is given. We present two methods for discrete-time linear systems which are based on complex approximation theory. We will also point out that one of the polynomial techniques can also be applied to certain continuous-time systems.

Keywords: parameter-dependent systems, polynomial approximation, ensemble reachability, infinite-dimensional systems, approximation
MSC: 30E10; 93B05; 93C05

1. EXTENDED ABSTRACT

Controlling ensembles of systems is motivated by a wide range of engineering applications. In robotics and systems engineering there has recently been much interest in studying motion control problems for spatio-temporal systems and infinite platoons of vehicles, cf. Bamieh et al. (2002), where control actions and measurements take place in a spatially distributed way. Also problems arising in quantum control (NMR spectroscopy) and the control of flocks falls into the area of ensemble control, cf. (Brockett, 2012, Section 2.4).

1.1 Problem Statement

In this work we investigate families of parameter-dependent linear control systems. First, we treat the discrete-time case, i.e. we consider

$$x_{t+1}(\theta) = A(\theta)x_t(\theta) + b(\theta)u_t, \tag{1}$$

where the matrices $A(\theta) \in \mathbb{C}^{n\times n}$ and the vectors $b(\theta) \in \mathbb{C}^n$ are assumed to depend continuously on the parameter $\theta \in \mathbf{P}$ which is varying over a nonempty compact set $\mathbf{P} \subset \mathbb{C}$ with empty interior. To make things not too technical, we assume that $\mathbf{P} \subset \mathbb{R}$ is a compact interval. To express these assumptions we will shortly write $(A, b) \in C_{n,n}(\mathbf{P}) \times C_n(\mathbf{P})$ in the following. We emphasize that the input u does *not* dependent on the parameter. Since we are interested in reachability properties we set the initial condition to zero, i.e. $x_0(\theta) = 0$ for every $\theta \in \mathbf{P}$. For $T > 0$ and $u = (u_0, ..., u_{T-1}) \in \mathbb{C}^{1\times T}$, let $\varphi(T, u)(\theta)$ denote the solution to (1), i.e.

$$\varphi(T, u)(\theta) = \sum_{\tau=0}^{T-1} A(\theta)^{T-1-\tau} b(\theta)u_\tau \,. \tag{2}$$

In ensemble control, the key point is that the input u has to be independent of the system parameter $\theta \in \mathbf{P}$ and a central question is the following reachability property: A pair (A, B) is called *uniformly ensemble reachable* if for any state $f \in C_n(\mathbf{P})$ and for any $\varepsilon > 0$ the are $T > 0$ and $u \in \mathbb{C}^{1\times T}$ such that

$$\|\varphi(T, u) - f\|_\infty = \sup_{\theta\in\mathbf{P}} \|\varphi(T, u)(\theta) - f(\theta)\| < \varepsilon. \tag{3}$$

Note that, it is an immediate consequence of the assumption that the input has to be independent of the parameter that exact reachability (i.e. $\varepsilon = 0$) is never possible.

The focus of this paper is to provide methods to compute suitable inputs for uniformly ensemble reachable pairs (A, b). That is, given a family of terminal states $f \in C_n(\mathbf{P})$ and a neighborhood $B_\varepsilon(f) = \{g \in C_n(\mathbf{P}) \,|\, \|f - g\|_\infty < \varepsilon\} \subset C_n(\mathbf{P})$ find $T > 0$ and an open-loop control $u \in \mathbb{C}^{1\times T}$ such that $\varphi(T, u) \in B_\varepsilon(f)$.

1.2 Known criteria for uniform ensemble reachability

In this section we recall relevant known results that prepare the ground to derive constructive methods for the computation of suitable inputs. We start with two necessary conditions, cf. (Dirr and Schönlein, 2021, Thm. 3): If the pair (A, b) is uniformly ensemble reachable, then

(N1) the pair $(A(\theta), b(\theta))$ is reachable for every $\theta \in \mathbf{P}$.
(N2) for any pair of distinct parameters $\theta, \theta' \in \mathbf{P}$, the spectra of $A(\theta)$ and $A(\theta')$ are disjoint:

$$\sigma(A(\theta)) \cap \sigma(A(\theta')) = \emptyset.$$

⋆ Research supported by the ERC project CHRiSHarMa DLV-682402.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

For single-input linear systems the following conditions are sufficient for uniform ensemble reachability, cf. (Dirr and Schönlein, 2021, Thm. 4 & Cor. 1).

Theorem 1. Let $\boldsymbol{P}$ be a compact interval. A pair (A, b) is uniformly ensemble reachable if it satisfies (N1) and (N2) and one of the following sufficiency conditions:

(S1) The characteristic polynomials of $A(\theta)$ take the form
$$z^n - (a_{n-1}z^{n-1} + \cdots + a_1 z + a_0(\theta))$$
for some $a_{n-1}, ..., a_1 \in \mathbb{C}$ and $a_0 \in C(\mathbf{P})$.

(S2) $A(\theta)$ has simple eigenvalues for each $\theta \in \boldsymbol{P}$.

We note that due to the condition (N2) the function a_0 in (S2) is necessarily injective. Thus, $a_0 \colon \mathbf{P} \to a_0(\mathbf{P})$ is one-to-one and onto and so $a_0(\mathbf{P})$ defines a Jordan arc. In certain cases, like the controlled harmonic oscillator, it happens that the sufficiency conditions (S1) and (S2) are satisfied at the same time. Also, if $\mathbf{P}$ is compact interval, it is shown in (Dirr and Schönlein, 2021, Thm. 3) that the conditions (S2) holds necessarily for an open and dense subset of $\mathbf{P}$.

For discrete-time single input systems the solution can also be written as
$$\begin{aligned}\varphi(T,u)(\theta) &= \sum_{\tau=0}^{T-1} u_\tau \big(A(\theta)\big)^{T-1-\tau}\, b(\theta) \\ &= \big(u_{T-1}I + \cdots + u_0 A(\theta)^{T-1}\big)\, b(\theta).\end{aligned}$$
Using the polynomial
$$p(z) = u_{T-1} + u_{T-2}\, z + \cdots + u_1\, z^{T-2} + u_0\, z^{T-1}$$
and it follows that (A, b) is uniformly ensemble reachable if and only if for every $f \in C_n(\mathbf{P})$ and for every $\varepsilon > 0$ there is a polynomial p such that
$$\|p(A)b - f\|_\infty < \varepsilon.$$
Thus, the construction of suitable inputs is equivalent to the construction of a suitable polynomial. In this work, for given $f \in C_n(\mathbf{P})$ and $\varepsilon > 0$, we will present sufficient conditions and methods for the construction of suitable inputs $u = (u_0, ..., u_{T-1}) \in \mathbb{C}^{1\times T}$. Since the inputs will be given in terms of the coefficients of an appropriate polynomial, it is required to determine the degree and the coefficients of the polynomial. Indeed, the degree of the polynomial determines the number of inputs that are required and the coefficients specify the input values.

2. APPROXIMATION THEORY

This section contains results from classical approximation theory that are used in the construction methods in Section 3. We start with the results due to Weierstrass. Recall that, given a function $f \colon [a, b] \to \mathbb{R}$ the nth Bernstein polynomials is given by
$$B_{f,n}(x) := \sum_{k=0}^{n} \binom{n}{k} f(a + \tfrac{k}{n}(b-a)) \left(\tfrac{x-a}{b-a}\right)^k \left(\tfrac{b-x}{b-a}\right)^{n-k}.$$
The following version of the classical Weierstrass Approximation Theorem has been shown by Gzyl and Palacios (1997).

Theorem 2. (Weierstrass first theorem).
Let $f \colon [a, b] \to \mathbb{R}$ satisfy a Lipschitz condition. Then, for $n \geq 3$ the sequence of Bernstein polynomials satisfies
$$\|f - B_{f,n}\|_\infty \leq \left(4\, M_f + \frac{(b-a)\, L_f}{2}\right) \frac{\sqrt{\log n}}{\sqrt{n}}.$$

Associated with a continuous function $f \colon \partial\mathbb{D} \to \mathbb{C}$ we consider the trigonometric polynomials due to Fejér, defined as
$$F_{f,n}(z) := \sum_{k=-n}^{n} \frac{n+1-|k|}{n+1} \hat{f}(k)\; z^k,$$
where
$$\hat{f}(k) := \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{is}) e^{-iks}\, \mathrm{d}s$$
denotes the kth *Fourier coefficient* of f. Similar to Gzyl and Palacios (1997) the following version of the trigonometric Weierstrass Approximation Theorem can be shown, cf. Natanson (1964).

Theorem 3. (Bernstein).
Suppose that $f \colon \partial\mathbb{D} \to \mathbb{C}$ satisfies a Lipschitz condition. Then, the sequence of trigonometric polynomials $(F_{f,n})_{n\in\mathbb{N}}$ converges uniformly to f and satisfies
$$\sup_{z\in\partial\mathbb{D}} |f(z) - F_{f,n}(z)| \leq 2\sqrt{2}\pi L_f \cdot \frac{\ln n}{n},$$
where $L_f > 0$ denotes the Lipschitz constant.

Moreover we make use of

Theorem 4. (Runge's little Theorem (1885)).
Let K be a compact subset in $\mathbb{C}$ such that $\mathbb{C} \setminus K$ is connected. If there is an open set Ω containing K such that f is holomorphic on Ω, then for every $\varepsilon > 0$ there is polynomial p such that
$$\sup_{z\in K} |f(z) - p(z)| < \varepsilon.$$

For Runge's little Theorem there are constructive procedures illustrated in the literature, cf. e.g. Remmert (1998). However, to the best of the authors knowledge, an explicit representation of the degree and the coefficients of the Runge polynomial in terms of the function f and $\varepsilon > 0$ was previously not available and is presented in Schönlein (2021).

3. COMPUTATIONAL METHODS FOR DISCRETE-TIME SYSTEMS

We take the sufficient conditions provided in Theorem 1 as a starting point and we provide for each case a constructive procedure to compute a suitable input for a given target function f and a given neighborhood $B_\varepsilon(f)$ of it. To this end, we will need the following notations. For $g \colon \Omega \subset \mathbb{C} \to \mathbb{C}$ we define
$$M_g := \max_{z\in\Omega} |g(z)|.$$
Also we say that g satisfies a Lipschitz condition, i.e. there exists a $L_g > 0$ such that
$$|g(z_1) - g(z_2)| \leq L_g\, |z_1 - z_2|$$
for all $z_1, z_2 \in \Omega$. We use $\mathrm{Lip}(\Omega)$ to denote the set of functions that satisfy a Lipschitz condition.

3.1 Method S1

The first method is based on sufficient condition (S1). Let $f \in \mathrm{C}_n(\mathbf{P})$ and $\varepsilon > 0$ be given. A suitable input is obtained by carrying out the following basic steps (details for step (A2) will be given below):

(A1) Compute the Jordan arc
$$a_0(\mathbf{P}) = \{a_0(\theta) | \, \theta \in \mathbf{P}\} \subset \mathbb{C}.$$
(A2) Compute the polynomials $p_1, ..., p_n$ such that
$$\sup_{z \in a_0(\mathbf{P})} |p_k(z) - f_k(a_0^{-1}(z))| < \varepsilon$$
for all $k = 1, ..., n$.
(A3) Set $g(z) := z^n - (a_{n-1}z^{n-1} + \cdots + a_1 z)$ and define
$$p(z) := \sum_{k=1}^{n} p_k\big(g(z)\big) z^{k-1}. \tag{4}$$

It remains to provide how to get the polynomials in step (A2). We distinguish the cases that $a_0(\mathbf{P})$ is real or complex. Also we will assume that
$$f_k \circ a_0^{-1} \in \mathrm{Lip}(a_0(\mathbf{P}))$$
for all $k = 1, ..., n$. The polynomials are determined via the constructive proofs to the approximation results stated in Section 2. More precisely:

(a1) If $a_0(\mathbf{P}) = [a, b] \subset \mathbb{R}$, then according to Theorem 2 we take p_k, $k = 1, ..., n$ as the Bernstein polynomials to the component functions $f_k \circ a_0^{-1}$ of degree n_k, where the degree $n_k \in \mathbb{N}$ is chosen such that
$$\sqrt{2}\left(4\, M_{f_k \circ a_0^{-1}} + \tfrac{(b-a)}{2}\, L_{f_k \circ a_0^{-1}}\right) \sqrt{\frac{\log n_k}{n_k}} \leq \varepsilon.$$
(a2) If $a_0(\mathbf{P}) \subset \partial\mathbb{D} = \{z \in \mathbb{C} \,|\, |z| = 1\}$, we first extend the functions $f_k \circ a_0^{-1}$ to $\partial\mathbb{D}$ by defining
$$f_k(z) := w_{k,1} + (w_{k,2} - w_{k,1}) \frac{z - z_{k,1}}{z_{k,2} - z_{k,1}}$$
for all $z \in \partial\mathbb{D} \setminus a_0(\mathbf{P})$, where $w_{k,1}$ and $w_{k,2}$ are the values of f_k at the end-points $z_{k,1}$ and $z_{k,2}$ of $a_0(\mathbf{P})$, respectively and take $p_k = F_{f_k \circ a_0^{-1}, n_k}$ with n_k such that
$$\left(2\sqrt{2}\pi L_{f_k \circ a_0^{-1}} \frac{\ln n_k}{n_k}\right) \leq \varepsilon.$$

Sufficient conditions such that Method S1 works are presented in the following result.

Theorem 5. Suppose (A, b) satisfies (N1), (N2) and (S1). Let $\varepsilon > 0$ and suppose that $f \in C_n(\mathbf{P})$ is such that $f_k \circ a_0^{-1} \in \mathrm{Lip}(a_0(\mathbf{P}))$ for all $k = 1, ..., n$ and assume that the Jordan arc $a_0(\mathbf{P})$ lies either on the real line or on the unit circle. Let
$$p(z) = p_0 + \cdots + p_k z^k$$
be the polynomial of degree k defined by (4). Then, at time $T - 1 = k \geq 3$ with inputs
$$u = (u_0, ..., u_{T-1}) = (p_k, ..., p_0)$$
one has
$$\|\varphi(T, u) - f\|_\infty < \varepsilon.$$

The construction yields that the degree of the polynomial defined in (4) is given by
$$\max_{k=1,...,n} n_k(k-1)n.$$

3.2 Method S2

The second method is based on the sufficient condition (S2). Let $f \in \mathrm{C}_n(\mathbf{P})$ and $\varepsilon > 0$ be given. After applying a change of coordinates $T(\theta)$ we consider the pair
$$T(\theta)^{-1} A(\theta) T(\theta) = \begin{pmatrix} a_1(\theta) & & \\ & \ddots & \\ & & a_n(\theta) \end{pmatrix}$$
$$T(\theta)^{-1} b(\theta) = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$
where $a_1, ..., a_n$ denote the distinct eigenvalue Jordan arcs. Let $\tilde{f}(\theta) = T(\theta)^{-1} f(\theta)$ and let $\|\cdot\|_M$ be a matrix norm that is submultiplicative to $\|\cdot\|$ and set
$$\|T\|_{M,\infty} := \sup_{\theta \in \mathbf{P}} \|T(\theta)\|_M.$$

A suitable input is obtained by carrying out the following basic steps (details for steps (B2) and (B3) will be given below):

(B1) Compute the Jordan arcs
$$a_k(\mathbf{P}) = \{a_k(\theta), | \, \theta \in \mathbf{P}\} \subset \mathbb{C}, \;\; k = 1, ..., n$$
and set $a(\mathbf{P}) = \cup_{k=1,...,n} a_k(\mathbf{P})$.
(B2) Compute the polynomials $p_1, ..., p_n$ such that
$$\sup_{z \in a_k(\mathbf{P})} |p_k(z) - \tilde{f}_k(a_k^{-1}(z))| < \tfrac{\varepsilon}{3\|T\|_{M,\infty}}$$
for all $k = 1, ..., n$.
(B3) Let $\alpha_{k,l} := \sup_{\theta \in \mathbf{P}} |p_k(a_l(\theta))|$ and define $h_k \colon a(\mathbf{P}) \to \mathbb{C}$ by
$$h_k(z) = \begin{cases} 1 & \text{if } z \in a_k(\mathbf{P}) \\ 0 & \text{if } z \in a(\mathbf{P}) \setminus a_k(\mathbf{P}) \end{cases}$$
for $k = 1, ..., n$. Compute via Runge's little Theorem the polynomials $q_1, ..., q_n$ such that
$$\sup_{z \in a(\mathbf{P})} |q_k(z) - h_k(z))| < \frac{\varepsilon}{3\|T\|_{M,\infty} \sum_{l=1}^{n} \alpha_{k,l}}$$
for all $k = 1, ..., n$.
(B4) Set
$$p(z) = \sum_{k=1}^{n} p_k(z) q_k(z). \tag{5}$$

The polynomials in step (B2) are obtained as follows. Again we distinguish the cases that the Jordan arcs are real or complex and assume that
$$\tilde{f}_k \circ a_k^{-1} \in \mathrm{Lip}(a_k(\mathbf{P})).$$

(b1) If $a_k(\mathbf{P}) = [a, b] \subset \mathbb{R}$ take p_k as the Bernstein polynomials corresponding to the component functions $\tilde{f}_k \circ a_k^{-1}$ of degree n_k, where the degree $n_k \in \mathbb{N}$ is chosen such that
$$\sqrt{2}\left(4\, M_{\tilde{f}_k \circ a_k^{-1}} + \tfrac{(b-a)}{2}\, L_{\tilde{f}_k \circ a_k^{-1}}\right) \sqrt{\tfrac{\log n_k}{n_k}} \leq \frac{\varepsilon}{3\|T\|_{M,\infty}}.$$
(b2) Let $a_k(\mathbf{P}) \subset \partial\mathbb{D} = \{z \in \mathbb{C} \,|\, |z| = 1\}$. Then, first extend the functions $\tilde{f}_k \circ a_k^{-1}$ to $\partial\mathbb{D}$ by defining
$$\tilde{f}_k(z) := w_{k,1} + (w_{k,2} - w_{k,1}) \frac{z - z_{k,1}}{z_{k,2} - z_{k,1}}$$
for all $z \in \partial\mathbb{D} \setminus a_0(\mathbf{P})$, where $w_{k,1}$ and $w_{k,2}$ are the values of $\tilde{f}_k$ at the end-points $z_{k,1}$ and $z_{k,2}$ of $a_0(\mathbf{P})$,

respectively and take $p_k = F_{\tilde{f}_k \circ a_0^{-1}, n_k}$ with n_k such that

$$\left(2\sqrt{2}\pi L_{\tilde{f}_k \circ a_0^{-1}} \frac{\ln n_k}{n_k}\right) \leq \frac{\varepsilon}{3\|T\|_{M,\infty}}.$$

The next result states sufficient conditions so that Method S2 yields an appropriate input sequence.

Theorem 6. Assume that (A, b) satisfies (N1), (N2) and (S2). Let $\varepsilon > 0$ and suppose that $f \in C_n(\mathbf{P})$ is such that $\tilde{f}_k \circ a_k^{-1} \in \mathrm{Lip}(a_k(\mathbf{P}))$ for all $k = 1, ..., n$ and assume that the Jordan arcs $a_k(\mathbf{P})$ are either on the real line or on the unit circle. Let

$$p(z) = p_0 + \cdots + p_k z^k$$

be the polynomial of degree k defined by (5). Then, at time $T - 1 = k \geq 3$ with inputs

$$u = (u_0, ..., u_{T-1}) = (p_k, ..., p_0)$$

one has

$$\|\varphi(T, u) - f\|_\infty < \varepsilon.$$

We close this section by noting that there are two ways how the Methods S1 and S2 can be extended to arbitrary Jordan arcs in the complex plane. First, using techniques from numerical conformal mapping theory together with approximation results due to Walsh, cf. Walsh (1965), Schönlein (2021). A second approach is to apply a mixture of open-loop and feedback control inputs of the form

$$u(t) + k(\theta)x_t(\theta).$$

This ansatz yields the possibility to enforce the sufficiency conditions (S1) or (S2). Indeed, due to the necessary condition (N1), the eigenvalues can be placed so that all Jordan arcs are on the real line or on the unit circle. For details we refer to Schönlein (2022).

4. CONTINUOUS-TIME SINGLE-INPUT SYSTEMS

In this section we sketch how the Method S2 can be also be applied to continuous-time single input systems

$$\frac{\partial x}{\partial t}(t, \theta) = A(\theta)x(t, \theta) + b(\theta)u(t) \tag{6}$$

that satisfy the conditions (N1), (N2) and (S2). Let $T(\theta)$ denote the change of coordinates as in Section 3.2. Then, for $u \in L^1([0, T], \mathbb{C}^m)$ the solution to (6) can be written as

$$\begin{aligned}\varphi(T, u)(\theta) &= \int_0^T \mathrm{e}^{(T-\tau)A(\theta)} b(\theta)u(\tau)\,\mathrm{d}\tau \\ &= T(\theta) \begin{pmatrix} \int_0^T u(\tau)\mathrm{e}^{(T-\tau)a_1(\theta)}\,\mathrm{d}\tau \\ \vdots \\ \int_0^T u(\tau)\mathrm{e}^{(T-\tau)a_n(\theta)}\,\mathrm{d}\tau \end{pmatrix}\end{aligned}$$

Note also that for continuous-time systems the time $T > 0$ can chosen arbitrarily, cf. (Dirr and Schönlein, 2021, Section 1). Thus, let $T > 0$ be fixed. We divide $[0, T]$ into $K \in \mathbb{N}$ intervals I_l, $l = 0, ..., K - 1$ of length $\tau > 0$ so that the mappings $\theta \mapsto \mathrm{e}^{\tau a_k(\theta)}$ are injective for all $k = 1, ..., n$. Then, we take piecewise constant input functions $u\colon [0, T] \to \mathbb{C}$ given by

$$u|_{I_l}(t) := u_l \in \mathbb{C} \tag{7}$$

for some complex numbers $u_0, ..., u_{K-1}$. The kth component of the solution is then

$$\varphi_k(T, u)(\theta) = \left(\frac{\mathrm{e}^{\tau a(\theta)} - 1}{\tau a(\theta)}\right) \sum_{l=0}^{K} \tau u_{K-l}\, \mathrm{e}^{l\tau a(\theta)}$$

Furthermore, it holds

$$\lim_{\tau \to 0} \frac{\mathrm{e}^{\tau z} - 1}{\tau z} = 1 \quad \text{for all } z \in \mathbb{C} \setminus \{0\}.$$

So, for any $\varepsilon > 0$ there is a $\tau^* > 0$ so that for any $\tau \in (0, \tau^*)$ we have

$$\left|\frac{\mathrm{e}^{\tau a_k(\theta)} - 1}{\tau a_k(\theta)} - 1\right| < \tfrac{\varepsilon}{2} \tag{8}$$

for any $a_k(\theta) \neq 0$, $k = 1, ..., n$. Let $K^* := \left\lfloor \frac{T}{\tau^*} \right\rfloor$, where $\lfloor x \rfloor = \max\{z \in \mathbb{Z} : z \leq x\}$. Thus, in terms of the polynomial

$$p(z) := \sum_{l=0}^{K} u_{K-l}\, z^l \tag{9}$$

the kth component of the solution satisfies

$$\left|\tau p(\mathrm{e}^{\tau a_k(\theta)}) - \varphi_k(1, u)(\theta)\right| < \tfrac{\varepsilon}{2} \tag{10}$$

for all $\tau \in (0, \tau^*)$ and all $k = 1, ..., n$. The significance of (10) is that it is independent of the input values $u_0, ..., u_{K-1}$. Thus, to compute suitable inputs values one can follow the steps of Method S2. For a detailed exposition we refer to Schönlein (2021).

REFERENCES

Bamieh, B., Paganini, F., and Dahleh, M.A. (2002). Distributed control of spatially invariant systems. *IEEE Transactions on Automatic Control*, 47(7), 1091–1107.

Brockett, R. (2012). Notes on the control of the Liouville equation. In F. Alabau-Boussouira, R. Brockett, O. Glass, J. LeRousseau, and E. Zuazua (eds.), *Control of Partial Differential Equations*, volume 2048 of *Lecture Notes in Mathematics*, pp. 101–129. Springer, Heidelberg.

Dirr, G. and Schönlein, M. (2021). Uniform and L^q-ensemble reachability of parameter-dependent linear systems. *Journal of Differential Equations*, 283, 216–262.

Gzyl, H. and Palacios, J.L. (1997). The Weierstrass approximation theorem and large deviations. *Am. Math. Mon.*, 104(7), 650–653.

Natanson, I. P. (1964). *Constructive Function Theory, Vol. I.* Ungar.

Remmert, R. (1998). *Classical Topics in Complex Function Theory.* Springer, New York.

Schönlein, M. (2021). Polynomial methods to construct inputs for uniformly ensemble reachable systems. *arXiv:2105.14963v2.*

Schönlein, M. (2022). Feedback equivalence and uniform ensemble reachability. *Linear Algebra and its Applications*, 646, 175–194.

Walsh, J. (1965). Interpolation and Approximation by Rational Functions in the Complex Domain. 4th ed. American Mathematical Society, Providence, R.I.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On Internal Stability of Diffusive-Coupling and the Dangers of Cancel Culture ★

Gal Barkai * **Leonid Mirkin** * **Daniel Zelazo** **

* *Faculty of Mechanical Engineering, Technion—IIT, Haifa 3200003, Israel*
(e-mails: galbarkai@campus.technion.ac.il & mirkin@technion.ac.il)
** *Faculty of Aerospace Engineering, Technion—IIT, Haifa 3200003, Israel*
(e-mail: dzelazo@technion.ac.il)

Abstract: We study internal stability in the context of diffusively-coupled control architecture, common in multi-agent systems (e.g. the celebrated consensus protocol). We derive a condition under which the system can be stabilized by no controller from that class. The condition says effectively that diffusively-coupled controllers cannot stabilize agents that share common unstable dynamics, directions included. This class always contains a group of homogeneous unstable agents, like integrators. We argue that the underlying reason is intrinsic cancellations of unstable agent dynamics by such controllers, even static ones, where directional properties play a key role. The intrinsic lack of internal stability explains the notorious behavior of some distributed control protocols when affected by measurement noise or exogenous disturbances.

Keywords: Multi-agent systems, structural properties, stability.

1. INTRODUCTION

A multi-agent system (MAS) is a collection of independent systems (agents) coupled via pursuit of a common goal. In large-scale MASs the information exchange between agents is normally limited to a subset of the agents, known as *neighbors*. Control laws using only information from neighboring agents are called *distributed*.

This work studies a class of distributed control laws, where only *relative* measurements are exchanged between neighbors. In other words, each agent has access only to the difference between its output and that of each of its neighbours. Relative sensing appears frequently in MAS tasks where absolute measurement are hard to obtain, such as space and aerial exploration and sensor localization, see (Smith and Hadaegh, 2005; Khan et al., 2009; Zelazo and Mesbahi, 2011b) and the references therein. Distributed control laws generated by relative information are called *diffusive*, and systems controlled by such laws are called *diffusively coupled*. Diffusive coupling appear naturally in consensus and synchronization problems (Olfati-Saber et al., 2007; Wieland et al., 2011), making them common in the MAS literature. However, diffusively-coupled systems behave poorly when affected by disturbances and noise. Measurement noise rapidly deteriorates performance (Zelazo and Mesbahi, 2011a, §III.A), and even dynamic controllers can hardly attenuate disturbances (Ding, 2015). To illustrate some of these traits, consider a simple example.

1.1 Motivating example

Reaching agreement between autonomous agents is a fundamental building block in multi-agent coordination (Ren and Beard, 2008). In its simplest form, it concerns a group of integrator agents described by $\dot{x}_i(t) = u_i(t)$, which need to synchronize their states x_i in a distributed manner. Namely, it is required to attain

$$\lim_{t\to\infty}\big(x_i(t) - x_j(t)\big) = 0, \quad \forall i, j. \tag{1}$$

by an appropriate choice of control signals u_i with access only to a states of neighboring agents, denoted by the set $\mathcal{N}_i$. This problem can be solved by the celebrated consensus protocol (Olfati-Saber et al., 2007), which is a diffusive state-feedback

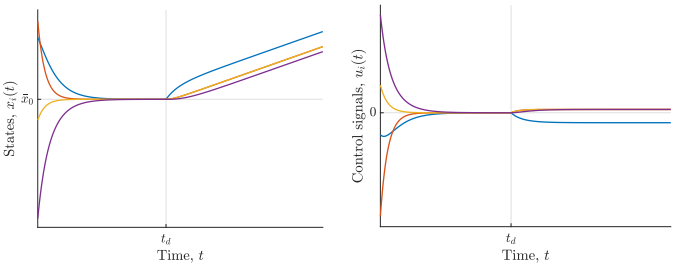
$$u_i(t) = -\sum_{j\in\mathcal{N}_i}\big(x_i(t) - x_j(t)\big). \tag{2}$$

If certain conditions on the communication topology hold, then the control law (2) drives the agents to agreement exponentially fast (Mesbahi and Egerstedt, 2010, Ch. 3).

This is no longer the case if the agent dynamics are affected also by exogenous inputs,

$$\dot{x}_i(t) = u_i(t) + d_i(t) \tag{3}$$

for some independent and unmeasurable load disturbances d_i. Fig. 1 demonstrates what happens with a group of 4 agents controlled by (2) when a unit step disturbance appears at one of them at some time instance $t = t_d$. For $t < t_d$, when the system is undisturbed, the states converge exponentially to the average



(a) states of the agents (b) control signals

Fig. 1. Consensus protocol for agents perturbed at $t = t_d$

★ Supported by the Israel Science Foundation (grants no. 2000/17 and 2285 /20).

of their initial conditions and the control signals go to zero. However, for $t > t_d$ the states x_i disagree and diverge, whereas the control signals u_i reach non-zero steady-state values.

The apparent instability of the whole system, manifested in the unboundedness of the states, can be explained by the well-known fact that the consensus protocol has a closed-loop eigenvalues at the origin (Olfati-Saber et al., 2007). Nevertheless, the boundedness of the control signals under such conditions is intriguing. The situation when some signal in the closed-loop system are bounded, whereas some other are not, may indicate unstable *pole-zero cancellations* in the feedback loop (Zhou et al., 1996, Ch. 5.3). However, controller (2) is static and thus has no zeros.

Still, the behavior like that in Fig. 1 prompts a deeper inspection of the *internal stability* property, which is the stability of all possible input/output relations in the system, see (Zhou et al., 1996; Skogestad and Postlethwaite, 2005). To the best of our knowledge, the instability phenomenon above was never explicitly connected with the lack of internal stability or unstable cancellations [1] . This is the starting point of the current study.

1.2 Contribution

In this note we show that the diffusive-coupling distributed control architecture for MASs is intrinsically internally unstable for many common agents configurations. Specifically, we prove that this is the case whenever all agents share common unstable dynamics (directions included for MIMO agents). This, for example, always happens in the group of homogeneous unstable agents, like integrator agents in (3).

We also explain the mechanism for the shown internal instability. It is indeed caused by unstable cancellations in the cascade of the block-diagonal aggregate plant and the diffusively coupled controller. Interesting is that these cancellations are caused not by controller zeros, but rather by an intrinsic spatial deficiency of the diffusive coupling configuration. They are thus independent of particular dynamics in processing relative measurements, only agents dynamics matter. It is worth mentioning that this instability mechanism is unrelated to the decentralized fixed modes (Wang and Davison, 1973).

The internal instability in the form of canceled plant poles explains then observed problems associated with the load disturbance response in some MAS applications.

Notations We extensively use standard notation from algebraic graph theory (Godsil and Royle, 2001). An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a finite vertex set $\mathcal{V}$ and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Denote by E the (oriented) incidence matrix of $\mathcal{G}$, defined component-wise by $[E(\mathcal{G})]_{ij} = 1$, when i is the head of edge j, $[E(\mathcal{G})]_{ij} = -1$ when i is the tail of edge j, and 0 otherwise. The matrix $L \coloneqq EE^\top$ is the combinatorial Laplacian matrix of $\mathcal{G}$. Note that $\mathbb{1} \in \ker E^\top$, thus L has an eigenvalue at the origin with $\mathbb{1}$ as its eigenvector.

The sets of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$ respectively, while the notations $\mathbb{C}_0$ and $\bar{\mathbb{C}}_0$ denote the open and closed right half complex plane, respectively. The complex-conjugate transpose of a matrix M is denoted by $M^\top$. The notation $\mathrm{diag}\{M_i\}$ stands for a block-diagonal matrix with diagonal elements M_i. The image (range) and kernel (null) spaces of a matrix M are notated $\mathrm{Im}\, M$ and $\ker M$, respectively. Given two matrices (vectors) M and N, $M \otimes N$ denotes their Kronecker product. By I_ν, or simply I, we denote the $\nu \times \nu$ identity matrix, by $\mathbb{1}_\nu$, or simply $\mathbb{1}$, the ν dimensional all-ones vector. The notation $\mathrm{spec}\, G$ refers to the set of eigenvalues if G is a matrix, or the set of poles if $G(s)$ is a proper transfer function. By H_∞ we denote the space of functions holomorphic and bounded in $\mathbb{C}_0$.

[1] Reminiscent reasoning has been mentioned in a formation control problem solved by a diffusive controller in (Fax and Murray, 2004, Sec. III.B), where the a cancelled mode was interpreted as unobservability of absolute motion.

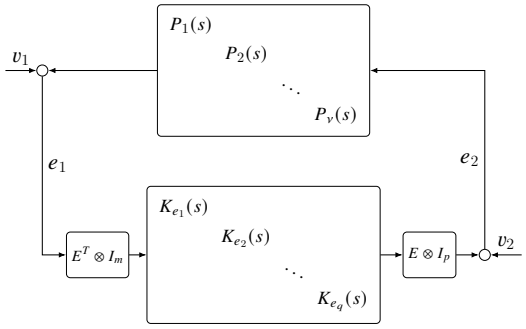


Fig. 2. Block diagram of the closed-loop

2. PROBLEM FORMULATION

Consider ν continuous-time agents P_i, each with m inputs and p outputs. Their aggregate dynamics are denoted as $P \coloneqq \mathrm{diag}\{P_i\}$, $i = 1, \ldots, \nu$. The interconnection topology of all agents is described by a graph $\mathcal{G}$ with ν nodes and q edges. Agent i and agent j are neighbors in the sense described in Section 1 if they are incident to the same edge. A dynamic controller, $K_e \coloneqq \mathrm{diag}\{K_{e,j}\}$, $j = 1, \ldots, q$, acts on the relative measurements on the edges. We assume hereafter that all P_i and K_{e_i} are linear time invariant (LTI), finite dimensional [2] , and that their transfer functions are proper.

A general diffusively-coupled architecture can be presented as the interconnection shown in Fig. 2, where the coupling matrix is the incidence matrix of $\mathcal{G}$. This representation is common in passivity-based analysis (Arcak, 2007), and sometimes called the canonical cooperative control structure (Sharf and Zelazo, 2017), (Bullo, 2022, Ch. 9). An equivalent representation can be made using the Laplacian matrix (Bullo, 2022, Ch. 8).

Note that the coupling matrices can be attached to either the plant or the controller, resulting in two distinct problems. One of them considers edge controllers stabilizing diagonal node dynamics (Bürger and De Persis, 2015), while the other a diagonal controller stabilizing the edge dynamics (Zelazo and Mesbahi, 2011a). In this note we consider the former, which includes controller

$$K \coloneqq (E \otimes I_m) K_e (E^\top \otimes I_p) \tag{4}$$

connected with a diagonal plant, as shown in Fig. 2, where v_1 and v_2 are arbitrary and bounded exogenous signals.

We say that the system in Fig. 2 is internally stable if all four closed-loop transfer functions connecting the exogenous signals v_1 and v_2 with the internal signals e_1 and e_2 are stable, i.e. belong to H_∞. The question studied in this note is under

[2] The arguments below could be extended to infinite-dimensional systems, but the involved technicalities are beyond the scope of this note.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

what conditions on the dynamics of the agents P_i there are edge controllers K_{e_j} internally stabilizing this system.

3. THE MAIN RESULT

The main result of this note provides a condition for the interconnection in Fig. 2 to be internally unstable irrespective of the choice of dynamics of K_e. As mentioned, the underlying reason is *cancellations* between poles of P and the controller K. It is well documented that poles cancelled in a cascade between a plant and controller are not truly cancelled (Anderson and Gevers, 1981). They may not appear in one I/O relation, but will still appear in a different one. For SISO systems cancellations are simple to spot and understand, they happen if and only if one system has a pole and the other a zero at the same location. When generalizing this to MIMO, poles and zeros have directional properties, see (Skogestad and Postlethwaite, 2005, §4.6.1) and (Mirkin, 2019, §3.4.2). Thus, directions, and not just locations, have to be considered.

Definition 1. Let G be an LTI system with m inputs and p outputs and (A, B, C, D) be its state space realization with realization poles at $\lambda_i \in \mathbb{C}$. The *input direction* of every λ_i is

$$\mathrm{pdir}_{\mathrm{i}}(G, \lambda_i) \coloneqq B^\top \ker([\lambda_i I - A]^\top) \subseteq \mathbb{C}^m$$

and its *output direction* is

$$\mathrm{pdir}_{\mathrm{o}}(G, \lambda_i) \coloneqq C \ker(\lambda_i I - A) \subseteq \mathbb{C}^p.$$

Pole directions span subspaces of either the input or output space. If λ_i is a hidden, i.e. uncontrollable or unobservable, mode of the realization (A, B, C, D), then both $\mathrm{pdir}_{\mathrm{i}}(G, \lambda_i) = \{0\}$ and $\mathrm{pdir}_{\mathrm{o}}(G, \lambda_i) = \{0\}$, which follows by PBH arguments. In the SISO case, $\mathrm{pdir}_{\mathrm{i}}(G, \lambda_i) = \mathrm{pdir}_{\mathrm{o}}(G, \lambda_i) = \mathbb{C}$ whenever λ_i is also a pole of the transfer function $G(s)$, i.e. every pole is excited by every input.

Cancelled poles correspond to unobservable (uncontrollable) modes in either KP or PK (Zhou et al., 1996, Thm. 5.7). The following Lemma provides conditions on pole directions of the agents in the diagram in Fig. 2 under which parts of the dynamics of agents are canceled by the controller K.

Lemma 2. Let P and K be as described in Section 2 and λ be a common pole of all agents P_i.

i) If
$$\bigcap_{i=1}^{\nu} \mathrm{pdir}_{\mathrm{o}}(P_i, \lambda) \neq \{0\},$$
then λ is an unobservable mode of KP.

ii) If
$$\bigcap_{i=1}^{\nu} \mathrm{pdir}_{\mathrm{i}}(P_i, \lambda) \neq \{0\},$$
then λ is an uncontrollable mode of PK.

The cancellation of λ in Lemma 2 is independent of the controller dynamics. In MIMO systems poles of the plant can be canceled not by zeros of the controller, but rather by a normal rank deficiency of the latter. This is exactly what happens in the diffusively-coupled interconnection in Fig. 2. Namely, the intrinsic singularity of the incidence matrix, present in every diffusive controller, might cancel plant poles. A formal condition for that is stated in Lemma 2.

Since cancelled poles remain poles of at least one closed-loop transfer function (Anderson and Gevers, 1981), the above Lemma immediately implies the main result.

Theorem 3. Let P_i, $i = 1, \ldots, \nu$, be LTI finite-dimensional agents with proper transfer functions. If $\lambda \in \bar{\mathbb{C}}_0$ is a pole of each one of them such that

$$\bigcap_{j=1}^{\nu} \mathrm{pdir}_{\mathrm{i}}(P_j, \lambda) \neq \{0\} \tag{5a}$$

or

$$\bigcap_{j=1}^{\nu} \mathrm{pdir}_{\mathrm{o}}(P_j, \lambda) \neq \{0\}, \tag{5b}$$

then the interconnection shown in Fig. 2 is internally unstable irrespective of the choice of K_e. Moreover, if this λ is not a zero of K_e, then condition (5a) implies that λ is the pole of the closed-loop transfer function from v_2 to e_1, while condition (5b) implies that λ is the pole of the closed-loop transfer function from v_1 to e_2.

Theorem 3 asserts that that any common dynamics, determined by poles and corresponding directions, are cancelled by the diffusive coupling. This has an interesting immediate corollary. If the agents are homogeneous they share their entire dynamics, both stable and unstable, thus the diffusive structure can be thought of as cancelling an *entire agent*. This not only proves the unobservability of the mode at the origin claimed in (Fax and Murray, 2004), but proves that every pole loses multiplicity in the cascade.

This may have ramifications not only about the stability of the system, but also of its maximal attainable performance. For example, it explains the observation reported in (Li et al., 2010), where the disturbance rejection performance measure of the entire system is upper bounded by that of a single, uncontrolled agent. It also generalizes the observation from (Zelazo and Mesbahi, 2011a), where it was shown that for integrator agents there is always an unobservable mode parallel to span $\mathbb{1}$. Since this direction is in the null space of the incidence matrix, noise or disturbances effecting this mode cannot be attenuated by a diffusive controller. Similarly, this cancellation explains why the cooperative disturbance rejection scheme of (Ding, 2015) cannot reject load disturbances, but only synchronize to it.

4. CONCLUDING REMARKS

We presented necessary conditions for internal stabilizability of diffusively-coupled LTI systems. In particular, we have shown that, for finite-dimensional agents, common dynamics are cancelled by the diffusive controller. The final conclusion is that in numerous multi-agent problems, one cannot simultaneously achieve a cooperative objective and guarantee internal stability using only relative measurements. Extending these results to time-varying graphs and more general systems are subject to current research.

REFERENCES

Anderson, B. and Gevers, M. (1981). On multivariable pole-zero cancellations and the stability of feedback systems. *IEEE Transactions on Circuits and Systems*, 28(8), 830–833.

Arcak, M. (2007). Passivity as a design tool for group coordination. *IEEE Trans. Automat. Control*, 52(8), 1380–1390.

Bullo, F. (2022). *Lectures on Network Systems*. Kindle Direct Publishing, 1.6 edition. URL `http://motion.me.ucsb.edu/book-lns`.

Bürger, M. and De Persis, C. (2015). Dynamic coupling design for nonlinear output agreement and time-varying flow control. *Automatica*, 51, 210–222.

Ding, Z. (2015). Consensus disturbance rejection with disturbance observers. *IEEE Transactions on Industrial Electronics*, 62(9), 5829–5837.

Fax, J.A. and Murray, R.M. (2004). Information flow and cooperative control of vehicle formations. *IEEE Trans. Automat. Control*, 49(9), 1465–1476.

Godsil, C.D. and Royle, G.F. (2001). *Algebraic Graph Theory*. Springer.

Khan, U.A., Kar, S., and Moura, J.M.F. (2009). Distributed sensor localization in random environments using minimal number of anchor nodes. *IEEE Transactions on Signal Processing*, 57(5), 2000–2016.

Li, Z., Duan, Z., Chen, G., and Huang, L. (2010). Consensus of multiagent systems and synchronization of complex networks: A unified viewpoint. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 57(1), 213–224.

Mesbahi, M. and Egerstedt, M. (2010). *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, Princeton.

Mirkin, L. (2019). Linear Control Systems. course notes, Faculty of Mechanical Eng., Technion—IIT. URL `http://leo.technion.ac.il/Courses/LCS/LCSnotes.pdf`.

Olfati-Saber, R., Fax, A., and Murray, R.M. (2007). Consensus and cooperation in networked multi-agent systems. *Proc. IEEE*, 95(1), 215–233.

Ren, W. and Beard, R.W. (2008). *Distributed Consensus in Multi-vehicle Cooperative Control: Theory and Applications*. Springer-Verlag, London.

Sharf, M. and Zelazo, D. (2017). A network optimization approach to cooperative control synthesis. *IEEE Control Syst. Lett.*, 1(1), 86–91.

Skogestad, S. and Postlethwaite, I. (2005). *Multivariable Feedback Control: Analysis and Design*. John Wiley & Sons, Chichester, 2nd edition.

Smith, R.S. and Hadaegh, F.Y. (2005). Control of deep-space formation-flying spacecraft; relative sensing and switched information. *Journal of Guidance, Control, and Dynamics*, 28(1), 106–114.

Wang, S.H. and Davison, E. (1973). On the stabilization of decentralized control systems. *IEEE Transactions on Automatic Control*, 18(5), 473–478.

Wieland, P., Sepulchre, R., and Allgöwer, F. (2011). An internal model principle is necessary and sufficient for linear output synchronization. *Automatica*, 47(5), 1068–1074.

Zelazo, D. and Mesbahi, M. (2011a). Edge agreement: Graph-theoretic performance bounds and passivity analysis. *IEEE Transactions on Automatic Control*, 56(3), 544–555.

Zelazo, D. and Mesbahi, M. (2011b). Graph-theoretic analysis and synthesis of relative sensing networks. *IEEE Transactions on Automatic Control*, 56(5), 971–982.

Zhou, K., Doyle, J.C., and Glover, K. (1996). *Robust and Optimal Control*. Prentice-Hall, Englewood Cliffs, NJ.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Towards a general port-Hamiltonian descriptor formalism for multi-phase flow dynamics

Harshit Bansal * **Wil Schilders** *

* *Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands, (e-mail: h.bansal@tue.nl, w.h.a.schilders@tue.nl).*

Abstract: Recently, port-Hamiltonian (pH) representations have been developed for multi-phase flow models, such as the Two-Fluid Model and the zero-slip Drift Flux Model (DFM), with non-quadratic Hamiltonian functionals, by eliminating constraints and writing a partial differential-algebraic system as a system with (only) partial differential equations. However, the existing multi-phase modelling framework is not modular enough since mathematical computations have to be performed again for even a small change, say a different governing equation of state, in the model description. Furthermore, a pH representation of the general DFM still does not exist, and the complicated, non-linear models may not always be amenable to the pH model formulation as per the current state-of-the-art. To this end, we make efforts towards developing a general pH descriptor formalism for non-linear multi-phase flow dynamics.

Keywords: port-Hamiltonian, descriptor realization, multi-phase, Drift Flux Model, non-linear

1. INTRODUCTION

Port-Hamiltonian (pH) representations for fluid flow models have been developed in de Wilde (2015); Bansal et al. (2021a); Bansal (2020); Bansal et al. (2021b). In the scope of two-phase flow models, the pH formulation of the Two-Fluid Model (TFM) and the special zero-slip case of the Drift Flux Model (DFM) have recently been developed; see Bansal et al. (2021a,b). However, as mentioned in Bansal et al. (2021a), for the general DFM, for e.g., the DFM with the Zuber-Findlay slip conditions, there have been technical challenges in obtaining a pH formalism.

In Bansal et al. (2021a,b), the pH formulation of multi-phase flow models has been obtained by eliminating constraints and writing a system of partial differential-algebraic equations (PDAEs) as a system with (only) partial differential equations (PDEs). As a consequence, the exisiting pH-based multi-phase modelling framework is not modular enough since mathematical computations have to be performed again for even a small change in the model description. In order to have a modular modelling framework, it is best to resort to pH descriptor realizations. Quite some work has been done in the area of pH descriptor formalism; see Beattie et al. (2018); Mehrmann and Morandin (2019); Mehrmann and Unger (2022) for instance, but the field is still in the initial stages of theoretical development, in particular for partial differential algebraic systems with non-quadratic Hamiltonian functionals, which is a representative feature of multi-phase flow dynamical models of interest. Furthermore, despite the developments, for complex physical systems in general, re-formulating the (nonlinear, coupled) PDAEs into a pH structure is very cumbersome and almost unrealizable. Moreover, based on some theoretical observations; see Sections 2.1 and 2.2, it seems that there is a scope of having a more generalized pH descriptor realization.

In addition, one of the key hindrances in obtaining a pH formulation of a model written in terms of (non-conservative) state variables could be due to the use of the standard $\mathcal{L}^2$ inner product; see Matignon and Helie (2013), wherein the authors have shown that the formal skew-adjointness of an operator holds if it is defined with respect to a weighted inner product. It is, hence, tempting to check if the operators, which are not formally skew-adjoint with respect to the $\mathcal{L}^2$ inner product, can be shown to be formally skew-adjoint with respect to a weighted one.

The main contributions of this paper are: *(i)* we showcase the limitations of the state-of-the-art pH descriptor model formulations, *(ii)* we propose a (general) novel pH descriptor formalism based on the notion of the weighted inner product, which requires solving homogeneous Sylvester equations, and test it on a simplified model problem.

The outline of the rest of this paper is as follows: In Section 2.1 and Section 2.2, we highlight the possible issues with the current state-of-the-art pH (descriptor) formalisms. In Section 2.3, we propose a new methodology to develop pH descriptor realizations of linear or non-linear problems. Section 3 deals with conclusions and future works.

2. TOWARDS A GENERAL PORT-HAMILTONIAN DESCRIPTOR REALIZATION

2.1 One possible methodology

We first recall definition of pH descriptor systems from Mehrmann and Unger (2022), which is a slight generalization of the one by Mehrmann and Morandin (2019).

Definition 1. A pH descriptor system is a system of differential(-algebraic) equations of the form

$$\mathbf{E}(t,\mathbf{x})\dot{\mathbf{x}}+\mathbf{r}(t,\mathbf{x}) = \mathbf{A}(t,\mathbf{x})\mathbf{z}(t,\mathbf{x})+\mathbf{B}(t,\mathbf{x})\mathbf{u},$$
$$\mathbf{y} = \mathbf{B}(t,\mathbf{x})^{\mathrm{T}}\mathbf{V}(t,\mathbf{x})^{\mathrm{T}}\mathbf{z}(t,\mathbf{x}),$$

with state $\mathbf{x}(t) \in \mathbb{R}^n$, input $\mathbf{u}(t) \in \mathbb{R}^m$, output $\mathbf{y}(t) \in \mathbb{R}^m$, the flow matrix $\mathbf{E}(t,\mathbf{x}) \in \mathbb{R}^{l\times n}$, time-flow function $\mathbf{r}(t,\mathbf{x}) \in \mathbb{R}^l$, effort function $\mathbf{z}(t,\mathbf{x}) \in \mathbb{R}^k$, structure-dissipation matrix $\mathbf{A} \in \mathbb{R}^{l\times k}$ with $\mathbf{VA} = \mathbf{J} - \mathbf{R}$, structure matrix $\mathbf{J} = -\mathbf{J}^{\mathrm{T}}$, dissipation matrix $\mathbf{R} = \mathbf{R}^{\mathrm{T}} \geq 0$, port matrix $\mathbf{B}(t,\mathbf{x}) \in \mathbb{R}^{l\times m}$, projection-type matrix $\mathbf{V}(t,\mathbf{x}) \in \mathbb{R}^{k\times l}$, and the gradient of the Hamiltonian $\mathcal{H}$ satisfies $\partial_{\mathbf{x}}\mathcal{H} = \mathbf{E}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{z}$, and $\partial_t\mathcal{H} = \mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{r}$ pointwise.

Remark 2. If $\mathbf{V} = \mathbf{I}$ and $\mathbf{z}(t,\mathbf{x}) \in \mathbb{R}^l$, then Definition 1 reduces to the one in Mehrmann and Morandin (2019).

For simplicity, in the example that follows, we will neglect dissipative effects. To point out the issues with the state-of-the-art descriptor formalism, let us consider a special form of the two-phase flow model as in Bansal et al. (2021a):

$$\begin{aligned}
\partial_t m_g &= -\partial_\xi(m_g v),\\
\partial_t m_\ell &= -\partial_\xi(m_\ell v),\\
\partial_t v &= -\partial_\xi\Big(\frac{v^2}{2}\Big) - \frac{1}{m_g+m_\ell}\partial_\xi p,\\
0 &= \frac{m_g}{\rho_g} + \frac{m_\ell}{\rho_\ell} - 1, \ \text{(volume conservation)},\\
0 &= p - \rho_g^2 u_g'(\rho_g), \ \text{(E.O.S. for gas phase)},\\
0 &= p - \rho_\ell^2 u_\ell'(\rho_\ell), \ \text{(E.O.S. for liquid phase)},
\end{aligned} \tag{1}$$

where $m_g := \alpha_g\rho_g$, $m_\ell := \alpha_\ell\rho_\ell$, α_ℓ and α_g, resp., denote liquid and gas void fraction, ρ_ℓ and ρ_g refer to the density of the liquid and the gas phase, resp., v is the common fluid velocity, p is the common pressure, $\xi \in \Omega$ denotes spatial domain, u_g and u_ℓ, resp., represent the internal energy of the gaseous and liquid phase, and u_i' with $i = \{g,\ell\}$ denote the partial derivative of the internal energy associated to the i-th phase w.r.t. the fluid density of the i-th phase.

The Hamiltonian functional for the above conservative zero-slip DFM can be expressed as:

$$\mathcal{H} = \int_\Omega \Big(m_g u_g(\rho_g) + m_\ell u_\ell(\rho_\ell) + \frac{1}{2}m_g v_g^2 + \frac{1}{2}m_\ell v_\ell^2\Big)d\Omega. \tag{2}$$

Remark 3. Consider that algebraic constraints in (1) *are eliminated and that we have three PDEs and as many number of unknowns. The model in this form admits a pH realization; see Bansal et al. (2021a). Hence, this simplified set up is a good model problem to check the potential of existing pH descriptor realizations and the need for generalization, if any. The vectors/matrices/operators, in accordance with Definition 1, are as follows:*

$$\mathbf{z} = \begin{pmatrix}\delta_{m_g}\mathcal{H}\\ \delta_{m_\ell}\mathcal{H}\\ \delta_v\mathcal{H}\end{pmatrix}, \ \mathbf{x} = \begin{pmatrix}m_g\\ m_\ell\\ v\end{pmatrix}, \ \mathbf{E} = \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & -\partial_\xi(\frac{m_g}{m_g+m_\ell}\cdot)\\ 0 & 0 & -\partial_\xi(\frac{m_\ell}{m_g+m_\ell}\cdot)\\ -\frac{m_g}{m_g+m_\ell}\partial_\xi & -\frac{m_\ell}{m_g+m_\ell}\partial_\xi & 0\end{pmatrix}.$$

We consider the mapping $\mathbf{V} = \mathbf{E}$. We notice that $\mathbf{VA} =: \mathbf{J}$ is skew-adjoint under periodic boundary conditions. We also have $\mathbf{E}^T\mathbf{V}^T\mathbf{z} = [\delta_{m_g}\mathcal{H}, \delta_{m_\ell}\mathcal{H}, \delta_v\mathcal{H}]^T = \delta_{\mathbf{x}}\mathcal{H}$.

Now consider that, unlike Remark 3, we do not eliminate constraints and $\mathbf{x} = [m_g, m_\ell, v, \rho_g, \rho_\ell, p]^T$. The variational derivatives are:

$$\begin{cases}\delta_{m_g}\mathcal{H} = \dfrac{v^2}{2} + u_g(\rho_g), \quad \delta_{m_\ell}\mathcal{H} = \dfrac{v^2}{2} + u_\ell(\rho_\ell),\\ \delta_v\mathcal{H} = (m_g+m_\ell)v, \quad \delta_{\rho_g}\mathcal{H} = m_g\dfrac{\partial u_g}{\partial\rho_g} = m_g u_g',\\ \delta_{\rho_\ell}\mathcal{H} = m_\ell\dfrac{\partial u_\ell}{\partial\rho_\ell} = m_\ell u_\ell', \quad \delta_p\mathcal{H} = 0, \ \text{and}\end{cases} \tag{3}$$

$$\mathbf{z} = \begin{pmatrix}\delta_{m_g}\mathcal{H}\\ \delta_{m_\ell}\mathcal{H}\\ \delta_v\mathcal{H}\\ \delta_{\rho_g}\mathcal{H}\\ \delta_{\rho_\ell}\mathcal{H}\\ \frac{m_g}{\rho_g}+\frac{m_\ell}{\rho_\ell}-1\\ p-\rho_g^2u_g'(\rho_g)\\ p-\rho_\ell^2u_\ell'(\rho_\ell)\end{pmatrix}, \mathbf{x} = \begin{pmatrix}m_g\\ m_\ell\\ v\\ \rho_g\\ \rho_\ell\\ p\end{pmatrix}, \mathbf{E} = \begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{pmatrix}.$$

As per Definition 1. we should have

$$\mathbf{E}^T\mathbf{V}^T\mathbf{z} = [\delta_{m_g}\mathcal{H}, \delta_{m_\ell}\mathcal{H}, \delta_v\mathcal{H}, \delta_{\rho_g}\mathcal{H}, \delta_{\rho_\ell}\mathcal{H}, 0]^T = \delta_{\mathbf{x}}\mathcal{H}.$$

However, any choice of $\mathbf{V} \in \mathbb{R}^{8\times6}$ along with other matrices will not be able to ensure $\mathbf{E}^T\mathbf{V}^T\mathbf{z} = \delta_{\mathbf{x}}\mathcal{H}$. We believe that a generalization to Definition 1 is hence possible. The aforementioned possibility of generalization holds irrespective of the structure of the operator A.

Remark 4. The aforementioned issue will exist even if we choose a different state vector, say $\mathbf{x} = [\alpha_g, \alpha_\ell, v, \rho_g, \rho_\ell, p]$.

Remark 5. We have observed that single-phase flow models can be cast in a pH descriptor form in line with Def. 1.

2.2 Another possible methodology

In order to explain more challenges, we next consider a simple setting in the scope of single-phase flow models. The model is governed by the following set of equations:

$$\begin{aligned}\partial_t\rho + \partial_\xi(\rho v) &= 0,\\ \partial_t(\rho v) + \partial_\xi(\rho v^2 + p) &= 0,\\ \rho = g(p) &= p/c^2,\end{aligned} \tag{4}$$

where ξ denotes the spatial domain, and c represents the speed of sound in the fluid medium. The Hamiltonian $\mathcal{H}$:

$$\mathcal{H}(\rho,v,p) = \int_\Omega \frac{1}{2}\rho v^2 + \rho\mathcal{U}(\rho)d\Omega, \ \text{where } d\mathcal{U} = \frac{p}{\rho^2}d\rho. \tag{5}$$

Two possible realizations (under the same choice of state coordinate vector $\mathbf{x} := [\rho, v, p]^T$) are as follows:

$$\underbrace{\begin{pmatrix}1&0&0\\0&1&0\\0&0&0\end{pmatrix}}_{\mathrm{E}}\begin{pmatrix}\partial_t\rho\\ \partial_t v\\ \partial_t p\end{pmatrix} = \underbrace{\begin{pmatrix}0&-1&0\\-1&0&0\\0&0&1\end{pmatrix}}_{\mathrm{J}}\partial_\xi\begin{pmatrix}\frac{\delta\mathcal{H}}{\delta\rho}\\ \frac{\delta\mathcal{H}}{\delta v}\\ \frac{\delta\mathcal{H}}{\delta p}\end{pmatrix} - \underbrace{\begin{pmatrix}0\\0\\ \frac{p}{\rho}g'(p)\end{pmatrix}}_{\mathrm{B}}.$$

$$\underbrace{\begin{pmatrix}1&0&0\\0&1&0\\1&0&-g'\end{pmatrix}}_{\mathrm{F}}\begin{pmatrix}\partial_t\rho\\ \partial_t v\\ \partial_t p\end{pmatrix} = \underbrace{\begin{pmatrix}0&-1&0\\-1&0&0\\0&0&0\end{pmatrix}}_{\tilde{J}}\partial_\xi\begin{pmatrix}\frac{\delta\mathcal{H}}{\delta\rho}\\ \frac{\delta\mathcal{H}}{\delta v}\\ \frac{\delta\mathcal{H}}{\delta p}\end{pmatrix}; \ g' = 1/c^2.$$

In view of above realizations, general representation reads:

$$\mathbf{E}\dot{\mathbf{x}} = \mathbf{J}\delta_{\mathbf{x}}\mathcal{H} - \mathbf{B}, \quad \text{where } \mathbf{E} \text{ is singular.} \tag{6}$$

or,

$$\mathbf{F}\dot{\mathbf{x}} = \tilde{\mathbf{J}}\delta_{\mathbf{x}}\mathcal{H}, \quad \text{where } \mathbf{F} \text{ is invertible.} \tag{7}$$

Building on (7), under choice of states: $\mathbf{x} := [\rho, v, p]^T$, we have: $\mathbf{F}\dot{\mathbf{x}} = \mathbf{J}\delta_{\mathbf{x}}\mathcal{H} = \mathbf{P_1}\partial_\xi\delta_{\mathbf{x}}\mathcal{H}$, which can be rewritten as:

$$\mathbf{F}\dot{\mathbf{x}} = \mathbf{P_1}\frac{\partial}{\partial\xi}(\frac{\delta\mathcal{H}}{\delta\mathbf{x}}). \tag{8}$$

The time-derivative of the Hamiltonian functional reads:

$$\frac{d}{dt}\mathcal{H} = \frac{\delta\mathcal{H}}{\delta\mathbf{x}}^T\dot{\mathbf{x}} = \frac{\delta\mathcal{H}}{\delta\mathbf{x}}^T\Big(\mathbf{F}^{-1}\mathbf{P_1}\frac{\partial}{\partial\xi}\Big)\frac{\delta\mathcal{H}}{\delta\mathbf{x}}. \tag{9}$$

The above equation can also be written as:

$$\frac{d}{dt}\mathcal{H} = \dot{\mathbf{x}}^T\frac{\delta\mathcal{H}}{\delta\mathbf{x}} = \frac{\delta\mathcal{H}}{\delta\mathbf{x}}^T\Big(\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi\Big)^T\frac{\delta\mathcal{H}}{\delta\mathbf{x}}.$$

Using the above two representations, the time derivative of the Hamiltonian functional can be written as:

$$\frac{d}{dt}\mathcal{H} = \frac{1}{2}\frac{\delta\mathcal{H}}{\delta\mathbf{x}}^T\underbrace{\Big(\mathbf{F}^{-1}\mathbf{P_1}\frac{\partial}{\partial\xi} + (\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi)^T\Big)}_{\mathbf{O}}\frac{\delta\mathcal{H}}{\delta\mathbf{x}}.$$

Analysis of the operator, $\mathbf{O}$, will dictate if we have the right properties of the matrices and the operators. $\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi + (\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi)^T = 0$ will enforce skew-adjointness and ensure that $\frac{d\mathcal{H}}{dt} = 0$ for a conservative system. We can write:

$$\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi + (\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi)^T = \mathbf{F}^{-1}\mathbf{P_1}\partial_\xi + (\mathbf{P_1}\partial_\xi)^T\mathbf{F}^{-T}. \tag{10}$$

The transpose of $\mathbf{P_1}\partial_\xi$ is: $(\mathbf{P_1}\partial_\xi)^T = -\mathbf{P_1}^T\partial_\xi$. A partial derivative acting on $\mathbf{F}^{-T}$ is known to be evaluated as:

$$\partial_\xi o\mathbf{F}^{-T} = \partial_\xi\mathbf{F}^{-T} + \mathbf{F}^{-T}\partial_\xi. \tag{11}$$

Using (11), the right-hand-side of (10) simplifies to:

$$\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi - \mathbf{P_1}^T\Big(\partial_\xi\mathbf{F}^{-T} + \mathbf{F}^{-T}\partial_\xi\Big). \tag{12}$$

Since $\mathbf{F}$ does not vary spatially, using (10) and (12), we have:

$$\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi + (\mathbf{P_1}\partial_\xi)^T\mathbf{F}^{-T} = \Big(\mathbf{F}^{-1}\mathbf{P_1} - \mathbf{P_1}^T\mathbf{F}^{-T}\Big)\partial_\xi. \tag{13}$$

For the representation given by (8) and $\rho = \frac{p}{c^2}$ (i.e., $g' = \frac{1}{c^2}$), the term $\Big(\mathbf{F}^{-1}\mathbf{P_1} - \mathbf{P_1}^T\mathbf{F}^{-T}\Big)$ turns out to be:

$$\Big(\mathbf{F}^{-1}\mathbf{P_1} - \mathbf{P_1}^T\mathbf{F}^{-T}\Big) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & c^2 \\ 0 & -c^2 & 0 \end{pmatrix}. \tag{14}$$

We would have wished to have the term at the right-hand-side of (14) to be zero. However, this is not the case. This hints that we need to introduce extra degrees of freedom, say through a weighting matrix, to ensure that we obtain a (pH) model formulation that has the right properties.

2.3 Proposed methodology

In Sections 2.1 and 2.2, we have seen that there exists a possibility to generalize existing pH model representations. We now propose a novel pH formalism; see Theorem 6.

Theorem 6. A mathematical model expressed in the form: $\mathbf{E}\dot{\mathbf{x}} = (\mathbf{J} - \mathbf{R})\mathbf{z}$*, where the inverse of the (known) matrix/operator* $\mathbf{E}$ *might not necessarily exist, possesses a pH structure if the gradient of the Hamiltonian satisfies* $\partial_{\mathbf{x}}\mathcal{H} = \mathbf{N}\mathbf{z}$*, the weighting matrix* $\mathbf{W}$ *is decomposed as* $\mathbf{W} = \mathbf{E}^\star\mathbf{M}\mathbf{E}$ *with* $\star$ *denoting the adjoint, the equation*

$$\Big(\mathbf{N}^\star\mathbf{E}^\star\mathbf{M}\mathbf{J} + \mathbf{J}^\star\mathbf{M}\mathbf{E}\mathbf{N}\Big) = 0 \tag{15}$$

holds, and $\Big(\mathbf{N}^\star\mathbf{E}^\star\mathbf{M}\mathbf{R} + \mathbf{R}^\star\mathbf{M}\mathbf{E}\mathbf{N}\Big)$ *is (formally) self-adjoint and positive semi-definite (PSD) for the matrix differential operator setting. Furthermore, if we ignore resistive effects, the operators* $\mathbf{E}, \mathbf{M}, \mathbf{J}, \mathbf{N}$ *should obey the relation:* $\Big(\mathbf{N}^\star\mathbf{E}^\star\mathbf{M}\mathbf{J} + \mathbf{J}^\star\mathbf{M}\mathbf{E}\mathbf{N}\Big) = 0$*. If we define* $\tilde{\mathbf{N}} = \mathbf{E}\mathbf{N}$*, then* (15)*, which can be written as:*

$$\Big(\tilde{\mathbf{N}}^\star\mathbf{M}\mathbf{J} + \mathbf{J}^\star\mathbf{M}\tilde{\mathbf{N}}\Big) = 0, \tag{16}$$

should hold. Here, $\mathbf{E}$ *and* $\mathbf{J}$ *are known quantities, and* $\mathbf{N}$ *and* $\mathbf{M}$ *are unknowns or degrees of freedom. For the conservative case, we could, in principle, make any choice for* $\mathbf{N}$ *(or* $\mathbf{M}$*) and find the (other) unknown* $\mathbf{M}$ *(or* $\mathbf{N}$*) by solving* (16)*. Inherently, the choice for* $\mathbf{N}$ *(or* $\mathbf{M}$*) should be such that the solutions to* (16) *exist. We could also assume that* $\tilde{\mathbf{N}}^\star\mathbf{M} = \mathbf{M}\tilde{\mathbf{N}} = \hat{\mathbf{M}}$*. Under this assumption, the model admits a pH structure if the following two relations hold:*

$$\hat{\mathbf{M}}\mathbf{J} + \mathbf{J}^\star\hat{\mathbf{M}} = 0, \quad \tilde{\mathbf{N}}^\star\mathbf{M} - \mathbf{M}\tilde{\mathbf{N}} = 0. \tag{17}$$

Proof. For the proof, we verify the conditions under which the rate of change of Hamiltonian, i.e., $\frac{d\mathcal{H}}{dt}$, along the solutions of the mathematical model is less than or equal to zero. To this end, we compute

$$\begin{aligned}\frac{d\mathcal{H}}{dt} &= \Big(\frac{\partial\mathcal{H}}{\partial\mathbf{x}}\Big)^\star\mathbf{W}\dot{\mathbf{x}} = \dot{\mathbf{x}}^\star\mathbf{W}\Big(\frac{\partial\mathcal{H}}{\partial\mathbf{x}}\Big) \\ &= \frac{1}{2}\Big(\Big(\frac{\partial\mathcal{H}}{\partial\mathbf{x}}\Big)^\star\mathbf{W}\dot{\mathbf{x}} + \dot{\mathbf{x}}^\star\mathbf{W}\Big(\frac{\partial\mathcal{H}}{\partial\mathbf{x}}\Big)\Big) \\ &= \frac{1}{2}\Big(\mathbf{z}^T\Big(\mathbf{N}^\star\mathbf{E}^\star\mathbf{M}\mathbf{J} + \mathbf{J}^\star\mathbf{M}\mathbf{E}\mathbf{N}\Big)\mathbf{z} - \\ &\qquad \mathbf{z}^T\Big(\mathbf{N}^\star\mathbf{E}^\star\mathbf{M}\mathbf{R} + \mathbf{R}^\star\mathbf{M}\mathbf{E}\mathbf{N}\Big)\mathbf{z}\Big).\end{aligned}$$

The dissipation inequality is satisfied (only) if the first term in the last equality equals zero, i.e., (15) holds, and the second term in the last equality, i.e., $\Big(\mathbf{N}^\star\mathbf{E}^\star\mathbf{M}\mathbf{R} + \mathbf{R}^\star\mathbf{M}\mathbf{E}\mathbf{N}\Big)$ is (formally) self-adjoint and PSD.

Remark 7. It can be shown that if $\mathbf{W} = \mathbf{E}^\star\mathbf{M}\mathbf{E} = \mathbf{I}$ *(identity matrix),* $\mathbf{E}\mathbf{E}^\star\mathbf{M} = \mathbf{I}$*, and* $\mathbf{N} = \mathbf{E}^\star$*, where* $\mathbf{E}$*, in Theorem 6, under restrictions on the dimension of the variable* $\mathbf{z}$*, carries the same meaning as in Definition 1, then Theorem 6 reduces to the definition in Mehrmann and Morandin (2019). Furthermore, if* $\mathbf{N}$*,* $\mathbf{E}$*, and* $\mathbf{J}$ *in Theorem 6, are resp. equivalent to* $\mathbf{E}^\star\mathbf{V}^\star$*,* $\mathbf{E}$*, and* $\mathbf{A}$ *(in absence of resistive effects) in Definition 1,* $\mathbf{W} = \mathbf{E}^\star\mathbf{M}\mathbf{E} = \mathbf{I}$*,* $\mathbf{E}\mathbf{E}^\star\mathbf{M} = \mathbf{I}$*, then Theorem 6, with no restriction on* $\mathbf{z}$*, reduces to the definition in Mehrmann and Unger (2022).*

Remark 8. The boundary effects have been ignored in the new formulation, but can be easily taken into account.

The first equation in (17) is an operator (homogeneous) Lyapunov/Riccati equation for a known $\mathbf{J}$. The second equation in (17) represents a homogeneous operator Sylvester equation for an unknown $\mathbf{M}$, and is similar to an operator T-Riccati equation for an unknown $\tilde{\mathbf{N}}$. To the best of our knowledge, no work exists in the scope of obtaining numerical solutions to the operator T-Riccati equations. However, it is worth mentioning that only recently some works as in Benner et al. (2022); Benner and Palitta (2020) have started to address the question pertaining to numerical solutions of (a general class of) matrix T-Riccati equations. As far as research on numerical solutions, exis-

tence, etc., of operator Riccati/Lyapunov/Sylvester equations is concerned, quite some work has been done, for e.g., see Curtain and Pritchard (1976).

In the sequel, we only consider the setting where one is interested in (non-linear) descriptor realizations of conservative physical system discussed in Section 2.2. Using the principles discussed earlier in Section 2.2 and Section 2.3, the rate of change of Hamiltonian $\mathcal{H}$ along the solutions of the model should be zero. Mathematically, this means:

$$\Big(\frac{\partial \mathcal{H}}{\partial \mathbf{x}}\Big)^T \Big(\mathbf{W}\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi + \Big(\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi\Big)^T \mathbf{W}\Big)\frac{\partial \mathcal{H}}{\partial \mathbf{x}} = 0, \quad (18)$$

where $\mathbf{F}$ and $\mathbf{P_1}$ carry the meaning as introduced in Section 2.2. For (18) to hold, we should have:

$$\mathbf{W}\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi + \Big(\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi\Big)^T \mathbf{W} = 0. \quad (19)$$

The above equation can be rewritten as shown next:

$$\begin{aligned} &\mathbf{W}\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi + \Big(\mathbf{P_1}\partial_\xi\Big)^T \mathbf{F}^{-T}\mathbf{W} = 0, \\ &\mathbf{W}\mathbf{F}^{-1}\mathbf{P_1}\partial_\xi - \mathbf{P_1}^T\partial_\xi \circ (\mathbf{F}^{-T}\mathbf{W}) = 0, \end{aligned} \quad (20)$$

where $\hat{a} \circ \hat{b}$ denotes the action of $\hat{a}$ on $\hat{b}$. Assuming that $\mathbf{F}$ and $\mathbf{P_1}$ are not spatially varying, (20) reduces to

$$\mathbf{W}\mathbf{F}^{-1}\mathbf{P_1} - \mathbf{P_1}^T\mathbf{F}^{-T}\mathbf{W} = 0, \quad (21)$$

which, for an unknown $\mathbf{W}$, is a homogeneous Sylvester equation. It is not trivial to solve such an equation using MATLAB or other tools since they return $\mathbf{W} = \mathbf{0}$. In order to obtain a non-zero solution, although non-unique, we next briefly discuss the solution methodology.

Remark 9. If $\mathbf{E} = \mathbf{I}$ *(identity matrix),* $\mathbf{N} = \mathbf{I}$*, and* $\mathbf{J}$ *is a (formally) skew-adjoint matrix differential operator, then* (16) *or* (17) *reduce to* (21) *for the representation* (8)*.*

Homogeneous Sylvester Equation Consider a simplified form of the term appearing on the left-hand-side of (21):

$$G(\mathbf{W}) = \mathbf{W}\tilde{\mathbf{A}} - \tilde{\mathbf{A}}^T\mathbf{W}, \quad (22)$$

where $\tilde{\mathbf{A}}$, in the sequel, denotes $\tilde{\mathbf{A}} = \mathbf{F}^{-1}\mathbf{P_1}$. Next, we form the unknown $\mathbf{W}$ in the following way:

$$\mathbf{W}_{eigenfn} = ab^T; \;\; a, b \in \mathbb{R}^n, \quad (23)$$

where n stands for the dimension of the (square) matrix/operator (say $\mathbf{J}$), a and b are eigenvectors associated to the matrix $\tilde{\mathbf{A}}^T$; i.e., $\tilde{\mathbf{A}}^T a = \lambda a$; $\tilde{\mathbf{A}}^T b = \mu b$ or $b^T\tilde{\mathbf{A}} = \mu b^T$. Next, $\mathbf{W}_{eigenfn}$ applied to $\tilde{\mathbf{A}}$ yields:

$$\mathbf{W}_{eigenfn}\tilde{\mathbf{A}} = ab^T\tilde{\mathbf{A}} = a\mu b^T = \mu\mathbf{W}_{eigenfn}, \quad (24)$$

and $\tilde{\mathbf{A}}^T$ applied to $\mathbf{W}_{eigenfn}$ yields:

$$\tilde{\mathbf{A}}^T\mathbf{W}_{eigenfn} = \tilde{\mathbf{A}}^T ab^T = \lambda ab^T = \lambda\mathbf{W}_{eigenfn}. \quad (25)$$

As a consequence, (22) can be written as:

$$G(\mathbf{W}_{eigenfn}) = (\lambda + \mu)\mathbf{W}_{eigenfn}. \quad (26)$$

It can be said that if both $\mathbf{W}_{eigenfn}$ and $\mathbf{W}^T_{eigenfn}$ are solutions to the homogeneous Sylvester equation, then $\mathbf{W}_{eigenfn} + \mathbf{W}^T_{eigenfn}$ is also its solution.

Using the above methodology, the weighting matrix $\mathbf{W}$ can be obtained for the problem considered in Section 2.2. As an example, for $c = 316$, we can compute the eigenvectors associated to $\mathbf{P_1}^T\mathbf{F}^{-T}$ and, subsequently, compute several possible values of $\mathbf{W}$ using (23). One possible solution is:

$$\mathbf{W} = \begin{bmatrix} 0.5 & 0.5 & 0 \\ -0.5 & -0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (27)$$

3. CONCLUSION AND FUTURE WORKS

This work attempted to develop a more general port-Hamiltonian (pH) descriptor framework for multi-phase fluid dynamical (and non-linear physical) systems. It is worth mentioning that although we have a methodology to obtain the weighting matrix $\mathbf{W}$, the matrix $\mathbf{W}$ is not invertible. The future work(s) will deal with computing an invertible $\mathbf{W}$, and with the identification of more conditions on different matrices/operators involved (e.g., $\mathbf{J}, \mathbf{R}$) such that numerical solutions to the resulting operator equations exist. We eventually aim to develop a (conditional) pH descriptor framework for the general case of the Drift Flux Model for which the state-of-the-art pH formalism (Definition 1) seems unsuitable.

ACKNOWLEDGEMENTS

We would like to acknowledge Hans Zwart and Philipp Schulze for insightful discussions.

REFERENCES

Bansal, H. (2020). *Structure-preserving model order reduction for drilling automation.* PhD thesis, Eindhoven University of Technology.

Bansal, H., Schulze, P., Abbasi, M., Zwart, H., Iapichino, L., Schilders, W., and van de Wouw, N. (2021a). Port-Hamiltonian formulation of two-phase flow models. *Systems & Control Letters*, 149, 104881.

Bansal, H., Zwart, H., Iapichino, L., Schilders, W., and van de Wouw, N. (2021b). Port-Hamiltonian modelling of fluid dynamics models with variable cross-section. *IFAC-PapersOnLine*, 54(9), 365–372. 24th International Symposium on Mathematical Theory of Networks and Systems MTNS 2020.

Beattie, C., Mehrmann, V., Xu, H., and Zwart, H. (2018). Linear port-Hamiltonian descriptor systems. *Mathematics of Control, Signals, and Systems*, 30(4), 17.

Benner, P., Iannazzo, B., Meini, B., and Palitta, D. (2022). Palindromic linearization and numerical solution of nonsymmetric algebraic T-Riccati equations. *BIT Numerical Mathematics.*

Benner, P. and Palitta, D. (2020). On the Solution of the Nonsymmetric T-Riccati Equation. *arXiv:2003.03693 [cs, math].* ArXiv: 2003.03693.

Curtain, R. and Pritchard, A.J. (1976). The infinite-dimensional riccati equation for systems defined by evolution operators. *SIAM Journal on Control and Optimization*, 14(5), 951–983.

de Wilde, H. (2015). *Port-Hamiltonian discretization of gas pipeline networks.* Master thesis, University of Groningen.

Matignon, D. and Helie, T. (2013). A class of damping models preserving eigenspaces for linear conservative port-Hamiltonian systems. *European Journal of Control*, 19.

Mehrmann, V. and Morandin, R. (2019). Structure-preserving discretization for port-Hamiltonian descriptor systems. *In Proceedings of the 59th IEEE Conference on Decision and Control*, 6663 – 6868.

Mehrmann, V. and Unger, B. (2022). Control of port-Hamiltonian differential-algebraic systems and applications. *arXiv:2201.06590 [cs, eess, math].* ArXiv: 2201.06590.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

An occupation kernel approach to optimal control

Rushikesh Kamalapurkar *

* *Assistant Professor, School of Mechanical Engineering, Oklahoma State University, Stillwater, OK 74078,* rushikesh.kamalapurkar@okstate.edu

Joel A. Rosenfeld **

** *Assistant Professor, Department of Mathematics and Statistics, University of South Florida, Tampa, Fl 33620,* rosenfeldj@usf.edu

Abstract: In this effort, a novel operator theoretic framework is developed for data-driven solution of optimal control problems. The developed methods focus on the use of trajectories (i.e., time-series) as the fundamental unit of data for the resolution of optimal control problems in dynamical systems. Trajectory information in the dynamical systems is embedded in a reproducing kernel Hilbert space (RKHS) through what are called occupation kernels. The occupation kernels are tied to the dynamics of the system through the densely defined Liouville operator. The pairing of Liouville operators and occupation kernels allows for lifting of nonlinear finite-dimensional optimal control problems into the space of infinite-dimensional linear programs over RKHSs.

Keywords: optimal control, operator theoretic methods in systems theory, nonlinear systems and control

1. INTRODUCTION

Numerical solutions of optimal control problems are obtained by using Pontryagin's maximum principle Pontryagin et al. (1962) to convert the optimal control problem into a two-point boundary value problem von Stryk and Bulirsch (1992); Betts (1998) or a nonlinear programming problem Hargraves and Paris (1987); Huntington (2007); Fahroo and Ross (2008); Rao et al. (2010); Darby et al. (2011); Garg et al. (2011). While there is a rich history of literature on the topic of numerical optimal control, the computational efficiency of numerical optimal control is limited by that of nonlinear programming, where solutions of large problems can be computationally prohibitive and the solutions, when available, are typically only locally optimal.

Based on the seminal work of Lasserre Lasserre (2010) on moments and positive polynomials, occupation measure approaches that convert a nonlinear optimal control problem into an infinite dimensional linear program that can be efficiently solved using sum of squares based convex programming methods were developed in results such as Lasserre et al. (2008); Majumdar et al. (2014); Claeys et al. (2016); Zhao et al. (2017).

⋆ This research was supported by the Air Force Office of Scientific Research (AFOSR) under contract numbers FA9550-20-1-0127 and FA9550-21-1-0134, and the National Science Foundation (NSF) under awards 2027976 and 2027999. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

While computationally efficient, techniques that utilize occupation measures are typically only applicable to systems where the functions that describe the dynamics, the cost functions, and the constraint sets are polynomials. The techniques developed in this paper also convert finite dimensional nonlinear optimal control problems into infinite dimensional linear programs, but utilize a reproducing kernel Hilbert space framework. An advantage of framing the infinite dimensional linear program within the reproducing kernel Hilbert space framework is that the developed tools are applicable to optimal control problems with a broader range of cost functions and constraining sets. Principally, the advantage is realized by exchanging the moment problem for occupation measures with the more flexible approximation abilities of reproducing kernel Hilbert spaces.

2. REPRODUCING KERNEL HILBERT SPACES

Definition 1. A real-valued *reproducing kernel Hilbert space* (RKHS), H, over a set $X \subset \mathbb{R}^n$ is a Hilbert space of functions $f : X \mapsto \mathbb{R}$ such that for every $x \in X$, the evaluation functional $E_x f := f(x)$ is bounded.

By the Reisz representation theorem, for each $x \in X$ there is a corresponding function $k_x \in H$ such that $\langle f, k_x \rangle_H = f(x)$, where $\langle f, g \rangle_H$ denotes the inner product. For each RKHS, there is a uniquely identified *kernel function*, $K(x, y) := \langle k_y, k_x \rangle_H$, such that for any finite collection of points, $\{x_i\}_{i=1}^M$, the corresponding Gram matrix, $(K(x_i, x_j))_{i,j=1}^M$, is positive semi-definite.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The importance of RKHSs lies in their ability to perform as function approximators. In particular, just as the collection of polynomials is dense inside of the space of continuous functions over compact subsets of $\mathbb{R}^n$, *universal* RKHSs are those spaces that are also dense in the space of continuous functions over compact subsets of $\mathbb{R}^n$. Moreover, the following lemma demonstrates that it is sufficient to consider linear combinations of the kernel functions themselves for function approximation when the kernel is in a universal RKHS (See (Steinwart and Christmann, 2008, Theorem 4.21)).

Lemma 1. Consider the subset $S := \{K(\cdot, y) : y \in X\}$ of a RKHS H over a set X with kernel K. Then span S is dense in H with respect to the Hilbert space norm. Moreover, if K is continuous, then span S is dense in H with respect to the uniform norm over restrictions to compact subsets of X.

3. PROBLEM FORMULATION

Let $H(Y)$ be a real-valued RKHS of continuous functions over the set Y. Let X and D be compact subsets of $\mathbb{R}^n$, U a compact subset of $\mathbb{R}^m$, $\Sigma := [0, T] \times X$, and $S = \Sigma \times U$. Throughout the rest of this manuscript the RKHSs $H(X)$, $H(D)$ and $H(\Sigma)$ denote the RKHSs obtained through the functions in $H(S)$ where the inputs have been projected to X, D, and Σ, respectively. Let $f : [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a locally Lipschitz function and consider the dynamical system

$$\dot{x} = f(t, x, u), \qquad x(0) = x_0 \in \mathbb{R}^n. \tag{1}$$

A state of the dynamical system corresponding to the initial condition x_0 and controller $u : [0, T] \mapsto \mathbb{R}^m$ will be written as $\phi_f(t; x_0, u)$.

For a fixed T, the optimal control problem is formulated as the need to minimize the cost

$$J(x(\cdot), u(\cdot)) = \int_0^T h(t, x(t), u(t))\mathrm{d}t + F(x(T)), \tag{2}$$

for functions $h \in H(S)$ and $F \in H(D)$, over the set of differentiable functions $x : [0, T] \to \mathbb{R}^n$ and continuous functions $u[0, T] \to \mathbb{R}^m$ subject to the constraints (1). For ease of exposition, the formulation considered here is more restrictive than strictly necessary. The methods developed in the following can be extended to include measurable control signals and absolutely continuous state trajectories.

In the following, occupation kernels and Liouville operators, first introduced in Rosenfeld et al. (2020) are utilized to lift the nonlinear optimal control problem into the space of infinite-dimensional linear programs.

4. OCCUPATION KERNELS AND THE COST FUNCTIONAL

Whenever $(t, x(t), u(t)) \in S$ for all $t \in [0, T]$, the functional $g \mapsto \int_0^T g(t, x(t), u(t))\mathrm{d}t$, that maps from $H(S)$ to $\mathbb{R}$, is linear and bounded. Indeed, given the kernel function K_S corresponding to $H(S)$, it can be seen that

$$\left|\int_0^T g(t,x(t),u(t))\mathrm{d}t\right| \leq \left|\int_0^T \langle g, K_S(\cdot,(t,x(t),u(t)))\rangle_{H(S)}\mathrm{d}t\right|$$

$$\leq T\|g\|_{H(S)} \sup_{[0,T]} \sqrt{K_S((t,x(t),u(t)),(t,x(t),u(t)))}$$

$$\leq T\|g\|_{H(S)} \sup_{y\in S} \sqrt{K_S(y,y)}.$$

As such, by the Reisz representation theorem, there exists a function $\Gamma_{x(\cdot),u(\cdot)} \in H(S)$ such that $\int_0^T g(t, x(t), u(t))\mathrm{d}t = \langle g, \Gamma_{x(\cdot),u(\cdot)}\rangle_{H(S)}$. The function $\Gamma_{x(\cdot),u(\cdot)}$ is the *occupation kernel* corresponding to the signals $x(\cdot)$ and $u(\cdot)$. Note that at this juncture, the signals $x(\cdot)$ and $u(\cdot)$ are independent, i.e., $x(\cdot)$ is not necessarily a trajectory of the dynamical system (1) in response to $u(\cdot)$.

The occupation kernel itself may be expressed as

$$\Gamma_{x(\cdot),u(\cdot)}(y) = \int_0^T K_S(y, (t, x(t), u(t)))\mathrm{d}t.$$

Moreover,

$$\|\Gamma_{x(\cdot),u(\cdot)}\|^2_{H(S)} = \int_0^T \int_0^T K((\tau,x(\tau),u(\tau)),(t,x(t),u(t)))d\tau\mathrm{d}t, \tag{3}$$

and when $K(x, y) = \Phi(\|x-y\|_2)$ is a radial basis function, such as the Wendland RBF or the Gaussian RBF, (3) may be bounded as $\|\Gamma_{x(\cdot),u(\cdot)}\|^2_{H(S)} \leq T^2\Phi(0)$.

Using the occupation kernels and the reproducing property $\langle F, K_D(\cdot, y)\rangle H_D = F(y)$ of the kernel function $K_D \in H_D$ corresponding to the RKHS $H(D)$, the cost functional in (2) can be expressed as

$$J(x(\cdot), u(\cdot)) = \langle h, \Gamma_{x(\cdot),u(\cdot)}\rangle_{H(S)} + \langle F, K_D(\cdot, x(T))\rangle_{H(D)}. \tag{4}$$

Note that the cost functional is linear with respect to the kernels $\Gamma_{x(\cdot),u(\cdot)}$ and K_D. If the dynamical system that constrains $x(\cdot)$ to be a solution in response to $u(\cdot)$ can also be expressed as a linear constraint on the space of kernels, the optimal control problem can be posed as a linear program in the infinite dimensional kernel space.

5. SYSTEM DYNAMICS AND THE TOTAL DERIVATIVE OPERATOR

In the following, a formulation of the dynamics in terms of total derivative operators is developed to construct the aforementioned linear constraint.

Definition 2. Define the *total derivative operator with symbol* f denoted by $A_f : \mathcal{D}(A_f) \to H(S)$ as $[A_f g](t, x, u) := \frac{\partial}{\partial t}g(t, x) + f(t, x, u) \cdot \nabla_x g(t, x)$ where the domain $\mathcal{D}(A_f)$ is defined cannonically as

$$\mathcal{D}(A_f) = \{g \in H(\Sigma) : A_f g \in H(S)\}. \tag{5}$$

The total derivative operator is seldom a compact operator. As such, to analyze the relationship between the total derivative operator and the occupation kernels, the theory of densely defined operators is leveraged.

Definition 3. (Densely Defined Operator). Given a set $\mathcal{D}(A) \subset H$, a linear operator $A : \mathcal{D}(A) \to H$ is said to be densely defined when $\mathcal{D}(A)$ is dense in H.

Differentiation is a canonical example of a densely defined operator. The following example, while not posed over a

RKHS, demonstrates this property of differentiation over the Hilbert space $L^2[0,1]$.

Example 1. Let $A = \frac{d}{dt}$ and suppose that the Hilbert space in question is $L^2[0,1]$. Since the derivative of any polynomial is again a polynomial and polynomials are dense in $L^2[0,1]$, $\mathcal{D}(A) := \{p : p \text{ is a polynomial over } [0,1]\}$ is a dense domain for A. It is also clear that $\mathcal{D}(A)$ cannot be extended to all of $L^2[0,1]$ as $f(t) = \sqrt{t}$ is in $L^2[0,1]$ and $\frac{d}{dt}f(t) = \frac{1}{2\sqrt{t}}$ is not.

The relationship between the total derivative operator and the occupation kernels is expressed through the adjoint of the total derivative operator, and for the development to be cogent, the adjoint needs to be densely defined. Since adjoints of closed operators over a Hilbert space are densely defined (Pedersen, 2012, Chapter 5), closedness of the total derivative operator is analyzed in the following.

Definition 4. Let A be an operator over H. A is said to be closed, if whenever $\{g_m\}_{m=1}^{\infty} \subset A$, $g_n \to f$ and $Ag_n \to h$ according to the Hilbert space norm, then $f \in \mathcal{D}(A)$ and $Af = h$.

The following theorem establishes a connection between the total derivative operator and signals $x(\cdot)$ and $u(\cdot)$ whenever $x(\cdot)$ is a solution of (1) under $u(\cdot)$. For brevity of notation, let $\Gamma_{x_0,u,f}$ denote the occupation kernel $\Gamma_{\phi_f(\cdot,x_0,u(\cdot)),u(\cdot)}$.

Theorem 1. The operator A_f introduced in Definition 2 is closed. Moreover, for an admissible trajectory $t \mapsto (t, x(t), u(t))$, with initial condition x_0, and that resides within a compact set for all $t \in [0,T]$, the function $\Gamma_{x_0,u,f}$ is in the domain of the adjoint of A_f.

Proof. Suppose that $\{g_m\}_{m=0}^{\infty} \subset \mathcal{D}(A_f) \subset H(\Sigma)$ such that $g_m \to g \in H(\Sigma)$ and $A_f g_m \to q \in H(S)$. Since the differentiability of the functions in H is inherited from the kernel function (see (Steinwart and Christmann, 2008, Corollary 4.36)), the function $\frac{\partial}{\partial x_i}g$ is well defined for each $g \in H(\Sigma)$ (but $\frac{\partial}{\partial x_i}g$ is not necessarily a function in $H(\Sigma)$). However, for any fixed t and x the mapping $p \mapsto \frac{\partial}{\partial x_i}p(t,x)$ is a continuous linear functional over $H(\Sigma)$. By (Steinwart and Christmann, 2008, Corollary 4.36),

$$\left|\frac{\partial}{\partial x_i}g_m(t,x) - \frac{\partial}{\partial x_i}g(t,x)\right| = \left|\frac{\partial}{\partial x_i}\left(g_m(t,x) - g(t,x)\right)\right| \leq \|g_m - g\|_{H(\Sigma)}\sqrt{\partial_i\partial_{i+n}K_\Sigma(((t,x)),((t,x)))}.$$

Hence, $\frac{\partial}{\partial x_i}g_m(t,x) \to \frac{\partial}{\partial x_i}g(t,x)$ for each $x \in X$ and $i = 1,\ldots,n$. Hence, $\frac{\partial}{\partial t}g_m(t,x) + f(t,x,u)\cdot\nabla_x g_m(t,x) \to \frac{\partial}{\partial t}g(t,x) + f(t,x,u)\cdot\nabla_x g(t,x)$ as $f(t,x,u)$ is constant with respect to m. Thus, $h = Ag$ and $g \in \mathcal{D}(A_f)$, and A_f is closed with the domain given in Definition 2.

To demonstrate that $\Gamma_{x_0,u,f}$ is in the domain of A_f^*, note that

$$\begin{aligned}&\left|\int_0^T \frac{\partial}{\partial t}g(t,x(t)) + f(t,x(t),u(t))\nabla_x g(t,x(t))dt\right| \\ &= \left|\int_0^T \dot{g}(t,x(t))dt\right| = |g(T,x(T)) - g(0,x(0))| \\ &= \left|\langle g, K_\Sigma(\cdot,(T,x(T))) - K_\Sigma(\cdot,(0,x(0)))\rangle_{H(\Sigma)}\right| \\ &\leq \|g\|_{H(\Sigma)}\|K_\Sigma(\cdot,(T,x(T))) - K_\Sigma(\cdot,(0,x(0)))\|_{H(\Sigma)}.\end{aligned}$$

Finally, given bounds on T and $\|x(t)\|_2$, a bound on $\|K_\Sigma(\cdot,(T,x(T))) - K_\Sigma(\cdot,(0,x(0)))\|_{H(\Sigma)}$ may be established. Thus, the functional over $\mathcal{D}(A_f)$ given as $g \mapsto \langle A_f g, \Gamma_{x_0,u,f}\rangle$ is bounded when $t \mapsto (t,x(t),u(t))$ is a trajectory of the system. It follows that the function $\Gamma_{x_0,u,f}$ is in the domain of the adjoint of the operator A_f. That is,

$$\langle A_f g, \Gamma_{x_0,u,f}\rangle_{H(S)} = \langle g, A_f^*\Gamma_{x_0,u,f}\rangle_{H(\Sigma)} = g(T,x(T)) - g(0,x(0)) \quad (6)$$

for all $g \in \mathcal{D}(A_f)$. □

Through consideration of (6) for an admissible trajectory satisfying the hypothesis of Theorem 1, $g \in \mathcal{D}(A_f)$ and setting $g_T(x) \equiv g(T,x) \in H(D)$, it can be observed that

$$\begin{aligned}\langle g, K_\Sigma(\cdot,(0,x_0))\rangle_{H(\Sigma)} &= g(0,x(0)) \\ &= -\langle g, A_f^*\Gamma_{x_0,u,f}\rangle + g(T,x(T)) \\ &= \langle -A_f g, \Gamma_{x_0,u,f}\rangle_{H(S)} + \langle g_T, K_D(\cdot,x(T))\rangle_{H(D)} \\ &= \langle(-A_f g, g_T), (\Gamma_{x_0,u,f}, K_D(\cdot,x(T)))\rangle_{H(S)\times H(D)}.\end{aligned}$$

Letting $\mathcal{L}_f : \mathcal{D}(A_f) \to H(S) \times H(D)$ denote the linear mapping $\mathcal{L}_f g = (-A_f g, g_T)$, it follows that $\langle g, \mathcal{L}_f^*(\Gamma_{x_0,u,f}, K_D(\cdot,x(T)))\rangle_{H(\Sigma)} = \langle g, K_\Sigma(\cdot,(0,x_0))\rangle_{H(\Sigma)}$ for all $g \in H(\Sigma)$. Hence, the linear constraint

$$\mathcal{L}_f^*(\Gamma_{x(\cdot),u(\cdot)}, K_D(\cdot,x(T))) = K_\Sigma(\cdot,(0,x_0)) \quad (7)$$

serves as a necessary condition for $x(\cdot)$ to be a trajectory of (1) in response to the control signal $u(\cdot)$.

6. A REFORMULATION OF THE OPTIMAL CONTROL PROBLEM

Using (4) and (7), the optimal control problem is expressed as an infinite dimensional linear program P given by

$$\min_{\Gamma_{x(\cdot),u(\cdot)},K_D(\cdot,x(T))} \langle(\Gamma_{x(\cdot),u(\cdot)},K_D(\cdot,x(T))),(h,F)\rangle_{H(S)\times H(D)}$$
$$\text{subject to: } \mathcal{L}_f^*(\Gamma_{x(\cdot),u(\cdot)},K_D(\cdot,x(T))) = K_\Sigma(\cdot,(0,x_0)).$$

To solve P, finite-dimensional representation of the decision variables $\Gamma_{x(\cdot),u(\cdot)}$ and $K_D(\cdot,x(T))$ is required. The representation is cogent under the following assumptions.

Assumption 1. A_f is densely defined on $H(\Sigma)$ together with a countable basis for $\mathcal{D}(A_f)$, given as $\{\sigma_m\}_{m=1}^{\infty} \subset \mathcal{D}(A_f)$. Furthermore, for all $s \in S$, the kernel functions satisfy $K_S(\cdot,s) \in \mathcal{D}(A_f)$.

Under Assumption 1, the optimal control problem can be expressed as the need to find the optimal real valued weights $\{w_i\}_{i=1}^{M_S}$ and $\{v_i\}_{i=1}^{M_D}$ that provide approximations for $\Gamma_{x(\cdot),u(\cdot)}$ and $K_D(\cdot,x(T))$ as

$$\Gamma_{x(\cdot),u(\cdot)}(\cdot) \approx \sum_{i=1}^{M_S} w_i K_S(\cdot,s_i) \quad (8)$$

$$K_D(\cdot,x(T)) \approx \sum_{i=1}^{M_D} v_i K_D(\cdot,d_i), \quad (9)$$

where $\{s_i\}_{i=1}^{M_S} \subset S$ is a collection of center in S, and $\{d_i\}_{i=1}^{M_D} \subset D$ is a collection of centers in D. The objective function of P can then be evaluated as

$$\langle(\Gamma_{x(\cdot),u(\cdot)}, K_D(\cdot, x(T))), (h, F)\rangle_{H(S)\times H(D)} \approx$$
$$\left\langle \left(\sum_{i=1}^{M_S} w_i K_S(\cdot, s_i), \sum_{i=1}^{M_D} v_i K_D(\cdot, d_i)\right), (h, F)\right\rangle_{H(S)\times H(D)}$$
$$= \sum_{i=1}^{M_S} w_i \langle K_S(\cdot, s_i), h\rangle_{H(s)} + \sum_{i=1}^{M_D} v_i \langle K_D(\cdot, d_i), F\rangle_{H(D)}$$
$$= \sum_{i=1}^{M_S} w_i h(s_i) + \sum_{i=1}^{M_D} v_i F(d_i). \quad (10)$$

Similarly, the constraint in P is satisfied provided $\langle \mathcal{L}_f g, (\Gamma_{x(\cdot),u(\cdot)}, K_D(\cdot, x(T)))\rangle_{H(S)\times H(D)} = g(0, x_0)$ for all $g \in \mathcal{D}(A_f)$, which in turn, is satisfied provided $\langle \mathcal{L}_f \sigma_m, (\Gamma_{x(\cdot),u(\cdot)}, K_D(\cdot, x(T)))\rangle_{H(S)\times H(D)} = \sigma_m(0, x_0)$ for all $m = 1, \cdots, \infty$. Selecting a finite set of basis functions $\{\sigma_1, \ldots, \sigma_{M_b}\}$, the constraint of P can thus be approximated using M_b linear constraints of the form

$$\sum_{i=1}^{M_D} v_i \sigma_m(T, d_i) - \sum_{i=1}^{M_S} w_i A_f \sigma_m(s_i) = \sigma_m(0, x_0), \quad (11)$$

for $m = 1, \cdots, M_b$. The optimal control problem thus admits the finite-rank representation

$$P_f : \min_{\{w_i\}_{i=1}^{M_S}, \{v_i\}_{i=1}^{M_D}} \sum_{i=1}^{M_S} w_i h(s_i) + \sum_{i=1}^{M_D} v_i F(d_i)$$
$$\text{subject to: } \sum_{i=1}^{M_D} v_i \sigma_m(T, d_i) - \sum_{i=1}^{M_S} w_i A_f \sigma_m(s_i) = \sigma_m(0, x_0),$$

for $m = 1, \cdots, M_b$. To ensure that the optimization problem is bounded, (3) may be employed as $\|\Gamma_{x_0,u,f}\|^2 \leq T^2\Phi(0)$, when K_S is the Gaussian or Wendland RBF, and $\|K(\cdot, x(T))\|^2 \leq \sup_{y\in D} K(y, y)$. Alternatively, $\Phi(0)$ may be replaced by an appropriate supremum bound. Depending on the selection of the kernel, a theoretically achievable approximation of $\Gamma_{x_0,u,f}$ and $K_D(\cdot, x(T))$ can be justified based on the density (or *fill distance*) of the centers within their respective parent sets.

7. CONCLUSION

In this abstract, the concepts of occupation kernels and total derivative operators are utilized to lift a nonlinear optimal control problem into a linear infinite-dimensional optimal control problem over functions in a RKHS. A finite-rank representation of the infinite-dimensional problem is obtained using kernel functions of the RKHSs and a countable basis for the domain of the total derivative operator. The authors plan to include an expanded introduction that places this work in the context of other lifting techniques such as occupation measures, provide a procedure to extract the optimal value function from a solution of P_f, and add a few example problems that demonstrate the utility of the developed methods.

REFERENCES

Betts, J.T. (1998). Survey of numerical methods for trajectory optimization. *J. Guid. Control Dynam.*, 21(2), 193–207.

Claeys, M., Daafouz, J., and Henrion, D. (2016). Modal occupation measures and LMI relaxations for nonlinear switched systems control. *Automatica*, 64, 143–154.

Darby, C.L., Hager, W.W., and Rao, A.V. (2011). An hp-adaptive pseudospectral method for solving optimal control problems. *Optim. Control Appl. Methods*, 32(4), 476–502. doi:10.1002/oca.957.

Fahroo, F. and Ross, I.M. (2008). Pseudospectral methods for infinite-horizon nonlinear optimal control problems. *J. Guid. Control Dynam.*, 31(4), 927–936.

Garg, D., Hager, W.W., and Rao, A.V. (2011). Pseudospectral methods for solving infinite-horizon optimal control problems. *Automatica*, 47(4), 829–837.

Hargraves, C.R. and Paris, S.W. (1987). Direct trajectory optimization using nonlinear programming and collocation. *J. Guid. Control Dynam.*, 10(4), 338–342.

Huntington, G.T. (2007). *Advancement and analysis of a Gauss pseudospectral transcription for optimal control*. Ph.D. thesis, Department of Aeronautics and Astronautics, MIT.

Lasserre, J.B., Henrion, D., Prieur, C., and Trélat, E. (2008). Nonlinear optimal control via occupation measures and LMI-relaxations. *SIAM J. Control Optim.*, 47(4), 1643–1666.

Lasserre, J.B. (2010). *Moments, Positive Polynomials and Their Applications*. Imperial College Press.

Majumdar, A., Vasudevan, R., Tobenkin, M.M., and Tedrake, R. (2014). Convex optimization of nonlinear feedback controllers via occupation measures. *Int. J. Robot. Res.*, 33(9), 1209–1230.

Pedersen, G.K. (2012). *Analysis now*, volume 118. Springer Science & Business Media.

Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., and Mishchenko, E.F. (1962). *The mathematical theory of optimal processes*. Interscience, New York.

Rao, A.V., Benson, D.A., Darby, C.L., Patterson, M.A., Francolin, C., and Huntington, G.T. (2010). Algorithm 902: GPOPS, a MATLAB software for solving multiple-phase optimal control problems using the Gauss pseudospectral method. *ACM Trans. Math. Softw.*, 37(2), 1–39.

Rosenfeld, J., Russo, B., Kamalapurkar, R., and Johnson, T. (2020). The occupation kernel method for nonlinear system identification. arXiv:1909.11792. Submitted to SIAM Journal on Control and Optimization.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Information Science and Statistics. Springer, New York.

von Stryk, O. and Bulirsch, R. (1992). Direct and indirect methods for trajectory optimization. *Ann. Oper. Res.*, 37(1), 357–373.

Zhao, P., Mohan, S., and Vasudevan, R. (2017). Control synthesis for nonlinear optimal control via convex relaxations. In *Proc. Am. Control Conf.*, 2654–2661.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Dirac structure for spatial multidimensional port-Hamiltonian Systems

Nathanael Skrepek *

* *Department of Mathematics and Science, University of Wuppertal, Germany, (e-mail: skrepek@uni-wuppertal.de)*

Abstract: We regard port-Hamiltonian systems on multidimensional spatial domains. We show that there are multiple (slightly different) Dirac structures assigned to such systems. Moreover, we point out that not every Dirac structure admits well-posedness of the corresponding system.

Keywords: Dirac structures, Port-Hamiltonian Systems, Infinite Dimensional Systems Theory.

1. INTRODUCTION

We regard the spatial multidimensional port-Hamiltonian systems and associate Dirac structures with theses systems. The Dirac structures that we investigate differ mainly in the boundary spaces. We will mainly focus on the linear port-Hamiltonian systems that were introduced in Skrepek (2021). However, we can also cover systems with a non quadratic Hamiltonians. In particular we associate a Dirac structure to systems of the form

$$\begin{aligned}\frac{\partial}{\partial t}x(t,\zeta) &= \sum_{i=1}^{n}\frac{\partial}{\partial \zeta_i}\begin{bmatrix}0 & L_i\\ L_i^{\mathsf{H}} & 0\end{bmatrix}\frac{\delta H(x(t,\zeta))}{\delta x} + P_0\frac{\delta H(x(t,\zeta))}{\delta x}\\ x(0,\zeta) &= x_0(\zeta),\end{aligned}$$

where L_i are matrices, P_0 is a skew-adjoint matrix, L_i^{H} denotes the Hermitian transposed (complex conjugated transposed) matrix of L_i, H is the Hamiltonian, $t \geq 0$ and $\zeta \in \Omega \subseteq \mathbb{R}^n$. Additionally, we will consider boundary ports later.

Dirac structures are one of the key elements in port-Hamiltonian modeling. They unify the description of complex interactions in physical systems. For further background see e.g. van der Schaft and Maschke (2002); Le Gorrec et al. (2005).

In Macchelli et al. (2004) they already introduced Dirac (Stokes-Dirac) structures for such systems, but they assume all functions to be smooth. Nevertheless they use the L^2 inner product as a dual pairing for the effort and flow space, which leads to an incomplete pairing. However, if you work with a complete L^2 space, the difficulty arises when you have to deal with trace operators. Also in Brugnoli et al. (2019) they regarded the Mindlin plate and showed that there is a Dirac structure associated to this system, but they refer to Macchelli et al. (2004) for the justification, which as already pointed out only deals with smooth function spaces.

The codomain of the trace operators will be *called boundary* space. One choice is $\mathsf{L}^2(\partial\Omega)$ as boundary space, which seems very natural but has disadvantages when it comes to solution theory. On the other hand we can consider the boundary spaces that were introduced in Skrepek (2021), which establish a *quasi Gelfand triple* with $\mathsf{L}^2(\partial\Omega)$ as pivot space. A quasi Gelfand triple is also a concept that was introduced in Skrepek (2021), which generalizes Gelfand triples (rigged Hilbert spaces). We will show for both choices of boundary spaces that there is Dirac structure associated to the system.

2. PRELIMINARY

Our general assumption is that $\Omega \subseteq \mathbb{R}^n$ is open and has a bounded Lipschitz boundary (Ω itself can be unbounded). Moreover, $\mathbb{K}$ denotes the scalar field, which can be either $\mathbb{R}$ or $\mathbb{C}$. We want to point out that the following is also possible with boundary operators that act only on a part of $\partial\Omega$. However, for simplicity we reduce ourselves to boundary operators that act on the entire boundary. For an extensive treatment of this section see Skrepek (2021).

Definition 2.1. Let $L = (L_i)_{i=1}^n$, where $L_i \in \mathbb{K}^{m_1\times m_2}$. Then we define

$$L_\partial \coloneqq \sum_{i=1}^{n}\partial_i L_i \quad \text{and} \quad L_\partial^{\mathsf{H}} \coloneqq (L^{\mathsf{H}})_\partial = \sum_{i=1}^{n}\partial_i L_i^{\mathsf{H}}$$

as operators from $\mathcal{D}'(\Omega)^{m_2}$ to $\mathcal{D}'(\Omega)^{m_1}$ and from $\mathcal{D}'(\Omega)^{m_1}$ to $\mathcal{D}'(\Omega)^{m_2}$, respectively (differential operators on the space of distributions). Furthermore, we define the space

$$\mathsf{H}(L_\partial,\Omega) \coloneqq \left\{f \in \mathsf{L}^2(\Omega,\mathbb{K}^{m_2}) \,\middle|\, L_\partial f \in \mathsf{L}^2(\Omega,\mathbb{K}^{m_1})\right\}.$$

This space is endowed with the inner product

$$\langle f,g\rangle_{\mathsf{H}(L_\partial,\Omega)} \coloneqq \langle f,g\rangle_{\mathsf{L}^2(\Omega,\mathbb{K}^{m_2})} + \langle L_\partial f, L_\partial g\rangle_{\mathsf{L}^2(\Omega,\mathbb{K}^{m_1})}.$$

We denote the outward pointing normed normal vector on $\partial\Omega$ by ν and its i-th component by ν_i. Moreover, we define

$$L_\nu \coloneqq \sum_{i=1}^{n}\nu_i L_i \colon \begin{cases}\mathsf{L}^2(\partial\Omega,\mathbb{K}^{m_2}) \to \mathsf{L}^2(\partial\Omega,\mathbb{K}^{m_1}),\\ f \mapsto \sum_{i=1}^{n}\nu_i L_i f,\end{cases}$$

and $L_\nu^{\mathsf{H}} \coloneqq (L^{\mathsf{H}})_\nu$.

We can and will regard L_∂ as an unbounded operator on $\mathsf{L}^2(\Omega)$ with domain $\mathsf{H}(L_\partial,\Omega)$.

For these differential operators exists an integration by parts formula, which is essentially a consequence of the generalized Stokes theorem. Since this integration by parts formula plays an important role for the Dirac structures we will introduce, these Dirac structure are also called

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Stokes-Dirac structures. We will first formulate the result for smooth functions. For a proof see (Skrepek, 2021, Lemma 3.8)

Lemma 2.2. For $f \in \mathsf{C}^\infty(\mathbb{R}^n)$ and $g \in \mathsf{C}^\infty(\mathbb{R}^n)$ we have

$$\langle L_\partial f, g\rangle_{\mathsf{L}^2(\Omega)} + \langle f, L_\partial^{\mathsf{H}} g\rangle_{\mathsf{L}^2(\Omega)} = \langle L_\nu f, g\rangle_{\mathsf{L}^2(\partial\Omega)}.$$

Clearly, we can easily extend this result by continuity to $f \in \mathsf{H}^1(\Omega)$ and $g \in \mathsf{H}^1(\Omega)$. However, we can even extend this result to $f \in \mathsf{H}(L_\partial, \Omega)$ and $g \in \mathsf{H}(L_\partial^{\mathsf{H}}, \Omega)$.

In order to do this we have to introduce new boundary spaces. First of all, since the inner product $\langle L_\nu f, g\rangle_{\mathsf{L}^2(\partial\Omega)}$ does only depend on the part of g which is in $\mathsf{L}^2_\pi(\partial\Omega) \coloneqq \overline{\operatorname{ran} L_\nu}$ we can add the orthogonal projection P_L from $\mathsf{L}^2(\partial\Omega)$ onto $\mathsf{L}^2_\pi(\partial\Omega)$ to g. Hence, we have

$$\langle L_\partial f, g\rangle_{\mathsf{L}^2(\Omega)} + \langle f, L_\partial^{\mathsf{H}} g\rangle_{\mathsf{L}^2(\Omega)} = \langle L_\nu f, P_L g\rangle_{\mathsf{L}^2(\partial\Omega)}.$$

To be more precise there is actually a trace operator involved in the $\mathsf{L}^2(\partial\Omega)$-inner product. We will denote the composition of this trace operator with P_L by π_L. We will endow $\operatorname{ran}\pi_L$ with the range norm

$$\|\phi\|_{\mathrm{ran}} \coloneqq \inf\left\{ \|g\|_{\mathsf{H}(L_\partial^{\mathsf{H}},\Omega)} \,\middle|\, \pi_L g = \phi, g \in \mathsf{C}(\mathbb{R}^n)\right\}$$

and denote the completion of this space (w.r.t. $\|\cdot\|_{\mathrm{ran}}$) by $\mathcal{V}_L$. There exists a continuous extension of π_L to $\mathsf{H}(L_\partial^{\mathsf{H}}, \Omega)$ denoted by $\bar\pi_L$. This extension is surjective. Moreover, we can also continuously extend L_ν to a surjective mapping from $\mathsf{H}(L_\partial, \Omega)$ to $\mathcal{V}_L'$ denoted by $\bar L_\nu$. Hence, we obtain

Corollary 2.3. For $f \in \mathsf{H}(L_\partial, \Omega)$ and $g \in \mathsf{H}(L_\partial^{\mathsf{H}}, \Omega)$ we have

$$\langle L_\partial f, g\rangle_{\mathsf{L}^2(\Omega)} + \langle f, L_\partial^{\mathsf{H}} g\rangle_{\mathsf{L}^2(\Omega)} = \langle \bar L_\nu f, \bar\pi_L g\rangle_{\mathcal{V}_L', \mathcal{V}_L}.$$

We say that $\bar L_\nu f$ is an element of $\mathsf{L}^2(\partial\Omega)$, if there is an $h_f \in \mathsf{L}^2(\partial\Omega)$ such that

$$\langle h_f, \pi_L g\rangle_{\mathsf{L}^2(\partial\Omega)} = \langle \bar L_\nu f, \pi_L g\rangle_{\mathcal{V}_L', \mathcal{V}_L} \quad \forall\, g \in \mathsf{C}^\infty(\mathbb{R}^n).$$

Accordingly, we say $\bar\pi_L g$ is in $\mathsf{L}^2(\partial\Omega)$, if there is an $h_g \in \mathsf{L}^2(\partial\Omega)$ such that

$$\langle L_\nu f, h_g\rangle_{\mathsf{L}^2(\partial\Omega)} = \langle L_\nu f, \bar\pi_L g\rangle_{\mathcal{V}_L', \mathcal{V}_L} \quad \forall\, f \in \mathsf{C}^\infty(\mathbb{R}^n).$$

3. PORT-HAMILTONIAN SYSTEMS

We regard port-Hamiltonian systems of the form

$$\begin{aligned}
\frac{\partial}{\partial t}x(t,\zeta) &= \sum_{i=1}^{n} \frac{\partial}{\partial \zeta_i}\begin{bmatrix} 0 & L_i \\ L_i^{\mathsf{H}} & 0\end{bmatrix}\frac{\delta H(x(t,\zeta))}{\delta x} \\
&\qquad + P_0 \frac{\delta H(x(t,\zeta))}{\delta x} \\
u(t,\xi) &= [0\ L_\nu]\,\frac{\delta H(x(t,\xi))}{\delta x} \\
y(t,\xi) &= [\pi_L\ 0]\,\frac{\delta H(x(t,\xi))}{\delta x} \\
x(0,\zeta) &= x_0(\zeta),
\end{aligned} \tag{1}$$

where $t \in \mathbb{R}_+$, $\zeta \in \Omega$, $\xi \in \partial\Omega$ and $\frac{\delta H(x)}{\delta x}$ denotes the variational derivative of the Hamiltonian H. The functions u and y are the boundary ports. We will regarding the state function $x\colon \mathbb{R}_+ \times \Omega \to \mathbb{K}^m$ as $x\colon \mathbb{R}_+ \to \mathsf{L}^2(\Omega;\mathbb{K}^m)$ by setting $x(t) = x(t,\cdot)$. This convention allows us to rewrite (1) as

$$\begin{aligned}
\dot x &= \left(\begin{bmatrix} 0 & L_\partial \\ L_\partial^{\mathsf{H}} & 0\end{bmatrix} + P_0\right)\frac{\delta H(x)}{\delta x}, \\
u &= \begin{bmatrix} 0 & \bar L_\nu\end{bmatrix}\frac{\delta H(x)}{\delta x}, \\
y &= [\bar\pi_L\ 0]\,\frac{\delta H(x)}{\delta x}, \\
x(0) &= x_0.
\end{aligned} \tag{2}$$

We regard this system in $\mathsf{L}^2(\Omega)$. In particular we say $-\dot x$ is the inner flow variable, $\frac{\delta H(x)}{\delta x}$ is the inner effort, u is the boundary flow and y is the boundary effort.

The following example will show that Maxwell's equations fit the previous system.

Example 3.1. Let us regard the matrices

$$L_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0\end{bmatrix},\ L_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0\end{bmatrix},\ \text{and } L_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0\end{bmatrix}.$$

We can easily see that $L_i^{\mathsf{H}} = -L_i$. Furthermore, the corresponding differential operator is

$$L_\partial = \begin{bmatrix} 0 & -\partial_3 & \partial_2 \\ \partial_3 & 0 & -\partial_1 \\ -\partial_2 & \partial_1 & 0\end{bmatrix} = \operatorname{rot} = -L_\partial^{\mathsf{H}}.$$

The corresponding operator L_ν that acts on $\mathsf{L}^2(\partial\Omega)$ can be written as a vector cross product

$$L_\nu f = \begin{bmatrix} 0 & -\nu_3 & \nu_2 \\ \nu_3 & 0 & -\nu_1 \\ -\nu_2 & \nu_1 & 0\end{bmatrix}\begin{bmatrix} f_1 \\ f_2 \\ f_3\end{bmatrix} = \nu \times f.$$

It can be calculated that the projection P_L on $\overline{\operatorname{ran} L_\nu}$ is given by

$$P_L g = (\nu \times g) \times \nu.$$

Hence, $\pi_L g = (\nu \times g\big|_{\partial\Omega}) \times \nu$.

Inserting the identities of the previous example into (2) gives a version of Maxwell's equations.

4. DIRAC STRUCTURES

We will regard the inner flow space $\mathcal{F}_{\mathrm{in}} \coloneqq \mathsf{L}^2(\Omega)$, the inner effort space $\mathcal{E}_{\mathrm{in}} \coloneqq \mathsf{L}^2(\Omega)$. For the boundary flow space $\mathcal{F}_\partial$ we will either use $\mathsf{L}^2_\pi(\partial\Omega)$, $\mathcal{V}_L'$, or $\{0\}$. Accordingly we will use for the boundary effort space $\mathcal{E}_\partial$ either $\mathsf{L}^2_\pi(\partial\Omega)$, $\mathcal{V}_L'$, or $\{0\}$. Clearly, we choose $\mathcal{F}_\partial$ and $\mathcal{E}_\partial$ always such that $\mathcal{F}_\partial' = \mathcal{E}_\partial$. We set flow space $\mathcal{F} \coloneqq \mathcal{F}_{\mathrm{in}} \times \mathcal{F}_\partial$ and effort space $\mathcal{E} \coloneqq \mathcal{E}_{\mathrm{in}} \times \mathcal{E}_\partial$.

Note that $\mathcal{F}$, $\mathcal{E}$ is a dual pair, its dual pairing is given, depending on the spaces $\mathcal{F}_\partial$ and $\mathcal{E}_\partial$, by

$$\begin{gathered}
\langle\cdot,\cdot\rangle_{\mathsf{L}^2(\Omega)} + \langle\cdot,\cdot\rangle_{\mathsf{L}^2(\partial\Omega)}, \\
\langle\cdot,\cdot\rangle_{\mathsf{L}^2(\Omega)} + \langle\cdot,\cdot\rangle_{\mathcal{V}_L',\mathcal{V}_L}, \\
\text{or } \langle\cdot,\cdot\rangle_{\mathsf{L}^2(\Omega)}.
\end{gathered}$$

Definition 4.1. Let $\mathcal{F}$, $\mathcal{E}$ be Banach spaces and $\langle\cdot,\cdot\rangle_{\mathcal{F},\mathcal{E}}$ a dual pairing for these spaces. Then we define the *canonical symmetric pairing*

$$\left\langle\!\!\left\langle \begin{bmatrix} f \\ e\end{bmatrix}, \begin{bmatrix} \hat f \\ \hat e\end{bmatrix}\right\rangle\!\!\right\rangle \coloneqq \langle f, \hat e\rangle_{\mathcal{F},\mathcal{E}} + \langle e, \hat f\rangle_{\mathcal{E},\mathcal{F}},$$

where $\langle e, \hat f\rangle_{\mathcal{E},\mathcal{F}} \coloneqq \overline{\langle \hat f, e\rangle_{\mathcal{F},\mathcal{E}}}$. We will denote orthogonality with respect to $\langle\!\langle\cdot,\cdot\rangle\!\rangle$ by $\perp_{\langle\!\langle,\rangle\!\rangle}$.

If $\mathcal{F} = \mathcal{E}$ is a Hilbert space, then the dual pairing $\langle \cdot, \cdot \rangle_{\mathcal{F},\mathcal{E}}$ is given by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$.

Definition 4.2. Let $\mathcal{F}$, $\mathcal{E}$ be a dual pair. Then $\mathcal{D} \subseteq \mathcal{F} \times \mathcal{E}$ is a *Dirac structure*, if $\mathcal{D}^{\perp_{\langle\!\langle\cdot,\cdot\rangle\!\rangle}} = \mathcal{D}$.

The space $\mathcal{F} \times \mathcal{E}$ is called the *bond space*.

Theorem 4.3. Let $\mathcal{F}_\partial = \mathcal{E}_\partial = \{0\}$. We will ignore the second component of $\mathcal{F}$ and $\mathcal{E}$ as $\mathcal{F} \cong \mathcal{F}_{\text{in}}$ and $\mathcal{E} \cong \mathcal{E}_{\text{in}}$. Then

$$\mathcal{D} \coloneqq \left\{ \begin{bmatrix} f \\ e \end{bmatrix} \in \mathcal{F} \times \mathcal{E} \;\middle|\; f = -\begin{bmatrix} 0 & L_\partial \\ L_\partial^{\mathsf{H}} & 0 \end{bmatrix} e,\ \begin{bmatrix} \bar{\pi}_L & 0 \end{bmatrix} e = 0 \right\}$$

is a Dirac structure.

We can also replace the condition $\begin{bmatrix} \bar{\pi}_L & 0 \end{bmatrix} e = 0$ in the previous theorem by $\begin{bmatrix} 0 & \bar{L}_\nu \end{bmatrix} e = 0$.

Proof. By Corollary 2.3, it is not hard to see that $J \coloneqq \left[\begin{smallmatrix} 0 & L_\partial \\ L_\partial^{\mathsf{H}} & 0 \end{smallmatrix}\right]$ with the boundary condition $\begin{bmatrix} \bar{\pi}_L & 0 \end{bmatrix} e = 0$ is skew-adjoint. Hence, for $\left[\begin{smallmatrix} \hat{f} \\ \hat{e} \end{smallmatrix}\right] \in \mathcal{D}^{\perp_{\langle\!\langle\cdot,\cdot\rangle\!\rangle}}$ the equation

$$\left\langle\!\!\left\langle \left[\begin{smallmatrix} -Je \\ e \end{smallmatrix}\right], \left[\begin{smallmatrix} \hat{f} \\ \hat{e} \end{smallmatrix}\right] \right\rangle\!\!\right\rangle = \langle -Je, \hat{e} \rangle_{\mathsf{L}^2(\Omega)} + \langle e, \hat{f} \rangle_{\mathsf{L}^2(\Omega)} = 0$$

implies $\hat{f} = -J\hat{e}$.

On the other by the skew-adjointness of J it follows immediately that $\left[\begin{smallmatrix} -Je \\ e \end{smallmatrix}\right] \perp_{\langle\!\langle\cdot,\cdot\rangle\!\rangle} \left[\begin{smallmatrix} -J\hat{e} \\ \hat{e} \end{smallmatrix}\right]$. □

Theorem 4.4. Let $\mathcal{F}_\partial = \mathcal{E}_\partial = \mathsf{L}^2_\pi(\partial\Omega)$. Then

$$\mathcal{D}_1 \coloneqq \left\{ \begin{bmatrix} f \\ f_\partial \\ e \\ e_\partial \end{bmatrix} \in \mathcal{F} \times \mathcal{E} \;\middle|\; f = -\underbrace{\begin{bmatrix} 0 & L_\partial \\ L_\partial^{\mathsf{H}} & 0 \end{bmatrix}}_{=:J} e, \begin{bmatrix} f_\partial \\ e_\partial \end{bmatrix} = \begin{bmatrix} 0 & \bar{L}_\nu \\ \bar{\pi}_L & 0 \end{bmatrix} e \right\}$$

is a Dirac structure.

Recall the canonical symmetric pairing is given by

$$\left\langle\!\!\left\langle \begin{bmatrix} f \\ f_\partial \\ e \\ e_\partial \end{bmatrix}, \begin{bmatrix} \hat{f} \\ \hat{f}_\partial \\ \hat{e} \\ \hat{e}_\partial \end{bmatrix} \right\rangle\!\!\right\rangle \coloneqq \langle f, \hat{e} \rangle_{\mathsf{L}^2(\Omega)} + \langle e, \hat{f} \rangle_{\mathsf{L}^2(\Omega)} + \langle f_\partial, \hat{e}_\partial \rangle_{\mathsf{L}^2(\partial\Omega)} + \langle e_\partial, \hat{f}_\partial \rangle_{\mathsf{L}^2(\partial\Omega)}.$$

Proof. It can be easily shown by Corollary 2.3 that $\mathcal{D}_1 \subseteq \mathcal{D}_1^{\perp_{\langle\!\langle\cdot,\cdot\rangle\!\rangle}}$.

Let $\begin{bmatrix} \hat{f} \\ \hat{f}_\partial \\ \hat{e} \\ \hat{e}_\partial \end{bmatrix} \in \mathcal{D}_1^{\perp_{\langle\!\langle\cdot,\cdot\rangle\!\rangle}}$. Then we have for any $\begin{bmatrix} f \\ f_\partial \\ e \\ e_\partial \end{bmatrix} \in \mathcal{D}_1$:

$$0 = \left\langle\!\!\left\langle \begin{bmatrix} f \\ f_\partial \\ e \\ e_\partial \end{bmatrix}, \begin{bmatrix} \hat{f} \\ \hat{f}_\partial \\ \hat{e} \\ \hat{e}_\partial \end{bmatrix} \right\rangle\!\!\right\rangle.$$

We choose $e \in \mathsf{C}_{\mathrm{c}}^\infty(\Omega)$, which implies that $\begin{bmatrix} -Je \\ 0 \\ e \\ 0 \end{bmatrix} \in \mathcal{D}_1$ and

$$0 = \left\langle\!\!\left\langle \begin{bmatrix} -Je \\ 0 \\ e \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{f} \\ \hat{f}_\partial \\ \hat{e} \\ \hat{e}_\partial \end{bmatrix} \right\rangle\!\!\right\rangle = \langle -Je, \hat{e} \rangle + \langle e, \hat{f} \rangle.$$

This gives

$$\langle Je, \hat{e} \rangle = \langle e, \hat{f} \rangle \quad \text{for all} \quad e \in \mathsf{C}_{\mathrm{c}}^\infty(\Omega),$$

and by distributional definition of J we conclude $\hat{f} = -J\hat{e}$. Note that according to the dimensions of the L_i we can split e into $\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$ and accordingly we can split $\hat{e}$. This gives

$$\begin{aligned} &\langle -Je, \hat{e} \rangle_{\mathsf{L}^2(\Omega)} + \langle e, -J\hat{e} \rangle_{\mathsf{L}^2(\Omega)} \\ &\quad = -\langle L_\partial e_2, \hat{e}_1 \rangle_{\mathsf{L}^2(\Omega)} - \langle L_\partial^{\mathsf{H}} e_1, \hat{e}_2 \rangle_{\mathsf{L}^2(\Omega)} \\ &\qquad - \langle e_1, L_\partial \hat{e}_2 \rangle_{\mathsf{L}^2(\Omega)} - \langle e_2, L_\partial^{\mathsf{H}} \hat{e}_1 \rangle_{\mathsf{L}^2(\Omega)} \\ &\quad = -\langle \bar{L}_\nu e_2, \bar{\pi}_L \hat{e}_1 \rangle_{\mathcal{V}_L', \mathcal{V}_L} - \langle \bar{\pi}_L e_1, \bar{L}_\nu \hat{e}_2 \rangle_{\mathcal{V}_L, \mathcal{V}_L'}. \end{aligned}$$

Hence, for $e \in \operatorname{dom} J \cap \{ g \mid \bar{\pi}_L g_1, \bar{L}_\nu g_2 \in \mathsf{L}^2_\pi(\partial\Omega) \}$ arbitrary we have

$$\begin{aligned} 0 &= \left\langle\!\!\left\langle \begin{bmatrix} -Je \\ \bar{L}_\nu e_2 \\ e \\ \bar{\pi}_L e_1 \end{bmatrix}, \begin{bmatrix} -J\hat{e} \\ \hat{f}_\partial \\ \hat{e} \\ \hat{e}_\partial \end{bmatrix} \right\rangle\!\!\right\rangle \\ &= -\langle \bar{L}_\nu e_2, \bar{\pi}_L \hat{e}_1 \rangle_{\mathcal{V}_L', \mathcal{V}_L} - \langle \bar{\pi}_L e_1, \bar{L}_\nu \hat{e}_2 \rangle_{\mathcal{V}_L, \mathcal{V}_L'} \\ &\qquad + \langle \bar{L}_\nu e_2, \hat{e}_\partial \rangle_{\mathsf{L}^2(\partial\Omega)} + \langle \bar{\pi}_L e_1, \hat{f}_\partial \rangle_{\mathsf{L}^2(\partial\Omega)}. \end{aligned}$$

Choose $e_1 \in \ker \bar{\pi}_L$ and $e_2 \in \mathsf{C}^\infty(\mathbb{R}^n)$ arbitrary, then

$$0 = \langle \bar{L}_\nu e_2, \hat{e}_\partial \rangle_{\mathsf{L}^2(\partial\Omega)} - \langle \bar{L}_\nu e_2, \bar{\pi}_L \hat{e}_1 \rangle_{\mathcal{V}_L', \mathcal{V}_L},$$

which implies $\hat{e}_\partial = \bar{\pi}_L \hat{e}_1$. Analogously, we can show that $\hat{f}_\partial = \bar{L}_\nu \hat{e}_2$, which implies $\begin{bmatrix} \hat{f} \\ \hat{f}_\partial \\ \hat{e} \\ \hat{e}_\partial \end{bmatrix} \in \mathcal{D}_1$ and therefore $\mathcal{D}_1^{\perp_{\langle\!\langle\cdot,\cdot\rangle\!\rangle}} \subseteq \mathcal{D}_1$. □

Theorem 4.5. Let $\mathcal{F}_\partial = \mathcal{V}_L'$ and $\mathcal{E}_\partial = \mathcal{V}_L$. Then

$$\mathcal{D}_2 \coloneqq \left\{ \begin{bmatrix} f \\ f_\partial \\ e \\ e_\partial \end{bmatrix} \in \mathcal{F} \times \mathcal{E} \;\middle|\; f = -\underbrace{\begin{bmatrix} 0 & L_\partial \\ L_\partial^{\mathsf{H}} & 0 \end{bmatrix}}_{=:J} e, \begin{bmatrix} f_\partial \\ e_\partial \end{bmatrix} = \begin{bmatrix} 0 & \bar{L}_\nu \\ \bar{\pi}_L & 0 \end{bmatrix} e \right\}$$

is a Dirac structure.

The proof is a copy of the proof of Theorem 4.4. Only in the last step (for $\hat{e}_\partial = \bar{\pi}_L \hat{e}_1$) we have to use the surjectivity of $\bar{L}_\nu$.

Note that the Dirac structures $\mathcal{D}_1$ from Theorem 4.4 and $\mathcal{D}_2$ from Theorem 4.5 are almost the same. In particular, we even have $\mathcal{D}_1 \subsetneq \mathcal{D}_2$ $(n > 1)$ and $\mathcal{D}_1$ is dense in $\mathcal{D}_2$. The reason why it is possible that nevertheless both are Dirac structures is because they are in different bond spaces.

For the initial boundary control systems the bond space of $\mathcal{D}_1$ is in some sense not suitable as the boundary operators do not map surjectively into $\mathsf{L}^2_\pi(\partial\Omega)$ (only densely). This is problematic for solution theory. On the other hand the bond space of $\mathcal{D}_2$ does not have this problem. And indeed it is possible to develop solution theory for this bond space, see Skrepek (2021).

5. CONCLUSION

One crucial tool to show that the defined sets are indeed Dirac structures were the boundary spaces $\mathcal{V}_L$ and $\mathcal{V}_L'$. These spaces helped us to make sure that all calculations are meaningful, even if we did not know whether the traces are $\mathsf{L}^2(\partial\Omega)$.

Moreover, with this general approach we have shown that there are Dirac structures for the wave equation, Maxwell's equations, and the Mindlin plate.

REFERENCES

Brugnoli, A., Alazard, D., Pommier-Budinger, V., and Matignon, D. (2019). Port-Hamiltonian formulation and symplectic discretization of plate models part i: Mindlin model for thick plates. *Applied Mathematical Modelling*, 75, 940–960. doi:10.1016/j.apm.2019.04.035.

Le Gorrec, Y., Zwart, H., and Maschke, B. (2005). Dirac structures and boundary control systems associated with skew-symmetric differential operators. *SIAM J. Control Optim.*, 44(5), 1864–1892. doi: 10.1137/040611677.

Macchelli, A., van der Schaft, A., and Melchiorri, C. (2004). Port Hamiltonian formulation of infinite dimensional systems i. modeling. In *2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601)*, volume 4, 3762–3767 Vol.4. doi: 10.1109/CDC.2004.1429324.

Skrepek, N. (2021). Well-posedness of linear first order port-Hamiltonian systems on multidimensional spatial domains. *Evol. Equ. Control Theory*, 10(4), 965–1006. doi:10.3934/eect.2020098.

van der Schaft, A.J. and Maschke, B.M. (2002). Hamiltonian formulation of distributed-parameter systems with boundary energy flow. *J. Geom. Phys.*, 42(1-2), 166–194. doi:10.1016/S0393-0440(01)00083-3.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A Distributed Algorithm for Measure-valued Optimization with Additive Objective ⋆

Iman Nodozi * **Abhishek Halder** **

* *Department of Electrical and Computer Engineering, University of California, Santa Cruz, CA 95064, USA (e-mail: inodozi@ucsc.edu).*
** *Department of Applied Mathematics, University of California, Santa Cruz, CA 95064, USA (e-mail: ahalder@ucsc.edu)*

Abstract: We propose a distributed nonparametric algorithm for solving measure-valued optimization problems with additive objectives. Such problems arise in several contexts in stochastic learning and control including Langevin sampling from an unnormalized prior, mean field neural network learning and Wasserstein gradient flows. The proposed algorithm comprises a two-layer alternating direction method of multipliers (ADMM). The outer-layer ADMM generalizes the Euclidean consensus ADMM to the Wasserstein consensus ADMM, and to its entropy-regularized version Sinkhorn consensus ADMM. The inner-layer ADMM turns out to be a specific instance of the standard Euclidean ADMM. The overall algorithm realizes operator splitting for gradient flows in the manifold of probability measures.

Keywords: Distributed algorithm, Wasserstein gradient flow, optimal transport.

1. INTRODUCTION

We consider measure-valued optimization problems of the form

$$\arg\inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} F_1(\mu) + F_2(\mu) + \ldots + F_n(\mu) \tag{1}$$

for some finite integer $n > 1$, where $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of Borel probability measures over $\mathbb{R}^d$ with finite second moments. We suppose that the functionals $F_i : \mathcal{P}_2(\mathbb{R}^d) \mapsto \mathbb{R}$ are convex for all $i \in [n]$. If the optimization in (1) is instead over $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, defined as the subset of $\mathcal{P}_2(\mathbb{R}^d)$ such that its elements are absolutely continuous w.r.t. the Lebesgue measure, then we can rewrite [1] (1) as

$$\arg\inf_{\rho} F_1(\rho) + F_2(\rho) + \ldots + F_n(\rho) \tag{2}$$

where the decision variable ρ is a joint PDF over $\mathbb{R}^d$ with finite second moment.

Problems of the form (1) and (2) arise in several contexts in statistics, machine learning, and control theory. This includes sampling from an unnormalized prior via Langevin Monte Carlo (see e.g., Stramer and Tweedie (1999a,b); Jarner and Hansen (2000); Roberts and Stramer (2002); Vempala and Wibisono (2019)), policy optimization in reinforcement learning (see e.g., Zhang et al. (2018); Chu et al. (2019); Zhang et al. (2020)), stochastic prediction (see e.g., Jordan et al. (1998); Ambrosio et al. (2005); Caluya and Halder (2019b,a)) and estimation (see e.g., Halder and Georgiou (2017, 2018, 2019)), density control (see e.g., Caluya and Halder (2021a,b)), mean field analysis of neural supervised (see e.g., Chizat and Bach (2018); Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Sirignano and Spiliopoulos (2020)) and unsupervised learning (see e.g., Domingo-Enrich et al. (2020)).

Let $F := F_1 + \ldots + F_n$. There is a natural connection between problems of the form (1) and that of the Wasserstein gradient flow

$$\frac{\partial\mu}{\partial t} = -\nabla^{W_2} F(\mu) := \nabla\cdot\left(\mu\frac{\delta F}{\delta\mu}\right), \tag{3}$$

where ∇ denotes the d dimensional Euclidean gradient, and $\frac{\delta}{\delta\mu}$ denotes the functional derivative w.r.t. μ. The operator ∇^{W_2} in (3) denotes the gradient w.r.t. the 2-Wasserstein metric W_2 between a pair of probability measures $\mu_x, \mu_y \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, defined as

$$W_2\left(\mu_x,\mu_y\right) := \left(\inf_{\pi\in\Pi\left(\mu_x,\mu_y\right)}\int_{\mathbb{R}^{2d}} c\left(\boldsymbol{x},\boldsymbol{y}\right)\,\mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y})\right)^{\frac{1}{2}}, \tag{4}$$

where $\Pi\left(\mu_x,\mu_y\right)$ is the set of joint probability measures or couplings over the product space $\mathbb{R}^{2d}$, having $\boldsymbol{x}$ marginal μ_x, and $\boldsymbol{y}$ marginal μ_y. We use the ground cost $c\left(\boldsymbol{x},\boldsymbol{y}\right) := \|\boldsymbol{x}-\boldsymbol{y}\|_2^2$, the squared Euclidean distance in $\mathbb{R}^d$. It is well-known (Villani, 2003, Ch. 7) that W_2 defines a metric on $\mathcal{P}_2\left(\mathbb{R}^d\right)$. For notational ease, we henceforth drop the subscript from W_2, and simply use W. The minimizer π^{opt} in (4) is referred to as the *optimal transportation plan*, and if $\mu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, then π^{opt} is supported on the graph of the *optimal transport map* T^{opt} pushing μ_x to μ_y.

The connection between (1) and (3) is that the minimizer of (1) may be realized as the stationary solution of (3). Conversely, if one is interested in the (possibly transient) solution of a PDE of the form (3), then it might be possible to compute the same by performing discrete time-stepping realizing gradient descent for (1).

⋆ This work is partially supported by NSF grants 1923278, 2112755.

[1] with slight abuse of notation in the sense (2) uses the same symbols F_i as in (1) for the additive functionals.

In recent years, several algorithms have been proposed for solving measure-valued optimization problems, see e.g., Benamou et al. (2016); Peyré (2015); Carlier et al. (2017); Carrillo et al. (2021); Mokrov et al. (2021); Alvarez-Melis et al. (2021). In this work, we explore the possibility of leveraging the additive structure of the objective in (1) for distributed nonparametric computation.

2. MAIN IDEA

We relabel the argument of the functional F_i in (1) as μ_i for all $i \in [n]$, and then impose the consensus constraint $\mu_1 = \mu_2 = \ldots = \mu_n$. Denoting

$$\mathcal{P}_2^{n+1}(\mathbb{R}^d) := \underbrace{\mathcal{P}_2(\mathbb{R}^d) \times \ldots \times \mathcal{P}_2(\mathbb{R}^d)}_{n+1 \text{ times}},$$

we then rewrite (1) as

$$\underset{(\mu_1,\ldots,\mu_n,\zeta)\in\mathcal{P}_2^{n+1}(\mathbb{R}^d)}{\arg\inf} \quad F_1(\mu_1) + F_2(\mu_2) + \ldots + F_n(\mu_n) \tag{5a}$$

$$\text{subject to} \quad \mu_i = \zeta \quad \text{for all } i \in [n]. \tag{5b}$$

Akin to the standard (Euclidean) augmented Lagrangian, we define the *Wasserstein augmented Lagrangian*

$$L_\alpha(\mu_1, \ldots, \mu_n, \zeta, \nu_1, \ldots, \nu_n) := \sum_{i=1}^{n} \left\{ F_i(\mu_i) + \frac{\alpha}{2} W^2(\mu_i, \zeta) + \int_{\mathbb{R}^d} \nu_i(\boldsymbol{\theta}) \left(\mathrm{d}\mu_i - \mathrm{d}\zeta\right) \right\} \tag{6}$$

where $\nu_i(\boldsymbol{\theta})$, $i \in [n]$, are the Lagrange multipliers for the constraints in (5b), and $\alpha > 0$ is a regularization constant.

Motivated by the Euclidean alternating direction method of multipliers (ADMM), we set up the recursions

$$\mu_i^{k+1} = \underset{\mu_i\in\mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; L_\alpha\left(\mu_1, \ldots, \mu_n, \zeta^k, \nu_1^k, \ldots, \nu_n^k\right) \tag{7a}$$

$$\zeta^{k+1} = \underset{\zeta\in\mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; L_\alpha\left(\mu_1^{k+1}, \ldots, \mu_n^{k+1}, \zeta, \nu_1^k, \ldots, \nu_n^k\right) \tag{7b}$$

$$\nu_i^{k+1} = \nu_i^k + \alpha\left(\mu_i^{k+1} - \zeta^{k+1}\right) \tag{7c}$$

where $i \in [n]$, and the recursion index $k \in \mathbb{N}_0$ (the set of whole numbers $\{0, 1, 2, \ldots\}$). We view (7a)-(7b) as primal updates, and (7c) as dual ascent.

Let $\nu_{\text{sum}}^k(\boldsymbol{\theta}) := \sum_{i=1}^{n} \nu_i^k(\boldsymbol{\theta})$, $k \in \mathbb{N}_0$. Substituting (6) in (7), dropping the terms independent of the decision variable in the respective arg inf, and re-scaling, the recursions (7) simplify to

$$\begin{aligned}\mu_i^{k+1} &= \underset{\mu_i\in\mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; \frac{1}{2} W^2\left(\mu_i, \zeta^k\right) + \frac{1}{\alpha}\left\{ F_i(\mu_i) + \int_{\mathbb{R}^d} \nu_i^k(\boldsymbol{\theta}) \mathrm{d}\mu_i \right\} \\ &= \mathrm{prox}^W_{\frac{1}{\alpha}\left(F_i(\cdot) + \int \nu_i^k \mathrm{d}(\cdot)\right)}\left(\zeta^k\right), \end{aligned} \tag{8a}$$

$$\begin{aligned}\zeta^{k+1} &= \underset{\zeta\in\mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; \sum_{i=1}^{n} \left\{ \frac{1}{2} W^2\left(\mu_i^{k+1}, \zeta\right) - \frac{1}{\alpha} \int_{\mathbb{R}^d} \nu_i^k(\boldsymbol{\theta}) \mathrm{d}\zeta \right\} \\ &= \underset{\zeta\in\mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; \left\{ \left(\sum_{i=1}^{n} W^2\left(\mu_i^{k+1}, \zeta\right) \right) - \frac{2}{\alpha} \int_{\mathbb{R}^d} \nu_{\text{sum}}^k(\boldsymbol{\theta}) \mathrm{d}\zeta \right\}, \end{aligned} \tag{8b}$$

$$\nu_i^{k+1} = \nu_i^k + \alpha\left(\mu_i^{k+1} - \zeta^{k+1}\right), \tag{8c}$$

wherein we use the notation $\mathrm{prox}^W_{G(\cdot)}(\zeta)$ to denote the *Wasserstein proximal operator* of the functional $G(\cdot)$, acting on $\zeta \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, given by

$$\mathrm{prox}^W_{G(\cdot)}(\zeta) := \underset{\mu\in\mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; \frac{1}{2} W^2(\mu, \zeta) + G(\mu). \tag{9}$$

We can view (9) as a generalization of the finite dimensional Euclidean proximal operator

$$\mathrm{prox}^{\|\cdot\|_2}_g(\boldsymbol{z}) := \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\inf} \; \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + g(\boldsymbol{x}). \tag{10}$$

We refer to (8) as the *Wasserstein consensus ADMM* – the notion generalizes its finite dimensional Euclidean counterpart in the sense (8a)-(8b) are analogues of the so-called x and z updates, respectively; see e.g., (Parikh and Boyd, 2014, Ch. 5.2.1). However, important difference arises in (8b) compared to its Euclidean counterpart due to the sum of squares of Wasserstein distances. In the Euclidean case, the corresponding z update can be analytically performed in terms of the arithmetic mean of the x updates. While (8b) does involve a *generalized mean* of the updates from (8a), we now have *Wasserstein barycentric proximal* of a linear functional. In other words, (8b) amounts to computing the Wasserstein barycenter of n measures $\{\mu_1^{k+1}, \ldots, \mu_n^{k+1}\}$ with a linear regularization involving ν_{sum}^k.

The proximal updates (8a) are closely related to the Wasserstein gradient flows generated by the respective (scaled) free energy functionals

$$\Phi_i(\mu_i) := F_i(\mu_i) + \int_{\mathbb{R}^d} \nu_i^k \mathrm{d}\mu_i, \quad \mu_i \in \mathcal{P}_2(\mathbb{R}^d), \quad i \in [n].$$

Under mild assumptions on Φ_i, as $1/\alpha \downarrow 0$, the sequence $\{\mu_i^k(\alpha)\}_{k\in\mathbb{N}_0}$ generated by the proximal updates (8a) converge to the measure-valued solution trajectory $\widetilde{\mu}_i(t, \cdot)$, $t \in [0, \infty)$, generated by the initial value problems (IVPs)

$$\frac{\partial \widetilde{\mu}_i}{\partial t} = -\nabla^W \Phi_i\left(\widetilde{\mu}_i\right), \; \widetilde{\mu}_i(t = 0, \cdot) = \widetilde{\mu}_i^0(\cdot), \; i \in [n]. \tag{11}$$

Thus, in a rather generic setting, performing the proximal updates (8a) in parallel across the index $i \in [n]$, amounts to performing distributed time updates for the approximate transient solutions of the IVPs (11).

Important examples of F_i include $\int V(\boldsymbol{\theta})\mathrm{d}\mu_i(\boldsymbol{\theta})$ (potential energy for some suitable advection potential V), $\beta^{-1}\int \log \mu_i(\boldsymbol{\theta})\mathrm{d}\mu_i(\boldsymbol{\theta})$ (internal energy with the "inverse temperature" parameter $\beta > 0$), $\int_{\mathbb{R}^{2d}} U(\boldsymbol{\theta}, \boldsymbol{\sigma})\mathrm{d}\mu_i(\boldsymbol{\theta})\mathrm{d}\mu_i(\boldsymbol{\sigma})$ (interaction energy for some symmetric positive definite interaction potential U).

To numerically realize the recursions (8), we consider a sequence of discrete probability distributions $\{\boldsymbol{\mu}_1^k, \ldots, \boldsymbol{\mu}_n^k, \boldsymbol{\zeta}^k\}$ indexed by $k \in \mathbb{N}_0$ where each distribution is a probability vector of length $N \times 1$, representative of the respective probability values at N samples. Thus, for each fixed $k \in \mathbb{N}_0$, the tuple

$$\left(\boldsymbol{\mu}_1^k, \ldots, \boldsymbol{\mu}_n^k, \boldsymbol{\zeta}^k\right) \in \underbrace{\Delta^{N-1} \times \ldots \times \Delta^{N-1}}_{n+1 \text{ times}} =: \left(\Delta^{N-1}\right)^{n+1}.$$

Likewise, for each fixed $k \in \mathbb{N}_0$, the Lagrange multipliers $\left(\boldsymbol{\nu}_1^k, \ldots, \boldsymbol{\nu}_n^k\right) \in \mathbb{R}^{nN}$, and $\boldsymbol{\nu}_{\text{sum}}^k = \sum_{i=1}^{n} \boldsymbol{\nu}_i^k \in \mathbb{R}^N$.

Given probability vectors $\boldsymbol{\xi}, \boldsymbol{\eta} \in \Delta^{N-1}$, let $\Pi_N\left(\boldsymbol{\xi}, \boldsymbol{\eta}\right) := \{\boldsymbol{M} \in \mathbb{R}^{N\times N} \mid \boldsymbol{M} \geq \boldsymbol{0} \text{ (elementwise)}, \boldsymbol{M}\boldsymbol{1} = \boldsymbol{\xi}, \boldsymbol{M}^\top\boldsymbol{1} = \boldsymbol{\eta}\}$. Also, let $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ denote the squared Euclidean distance matrix for the sampled data $\{\boldsymbol{\theta}_r\}_{r\in[N]}$ in $\mathbb{R}^d$, i.e.,

the entries of the matrix $\boldsymbol{C}$ are $\boldsymbol{C}(i,j) := \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2$ for all $i,j \in [N]$.

For each $i \in [n]$ and $k \in \mathbb{N}_0$, we write the discrete version of (8) as

$$\begin{aligned}\boldsymbol{\mu}_i^{k+1} &= \operatorname{prox}^{W}_{\frac{1}{\alpha}\left(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle\right)}\left(\boldsymbol{\zeta}^k\right)\\ &= \operatorname*{arg\,inf}_{\boldsymbol{\mu}_i \in \Delta^{N-1}} \left\{ \min_{\boldsymbol{M}\in\Pi_N(\boldsymbol{\mu}_i,\boldsymbol{\zeta}^k)} \frac{1}{2}\langle \boldsymbol{C}, \boldsymbol{M}\rangle + \frac{1}{\alpha}\left(F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle\right)\right\},\end{aligned} \tag{12a}$$

$$\boldsymbol{\zeta}^{k+1} = \operatorname*{arg\,inf}_{\boldsymbol{\zeta}\in\Delta^{N-1}} \left\{ \left(\sum_{i=1}^{n} \min_{\boldsymbol{M}_i \in \Pi_N\left(\boldsymbol{\mu}_i^{k+1}, \boldsymbol{\zeta}\right)} \frac{1}{2}\langle \boldsymbol{C}, \boldsymbol{M}_i\rangle \right) - \frac{2}{\alpha}\langle \boldsymbol{\nu}_{\text{sum}}^k, \boldsymbol{\zeta}\rangle \right\}, \tag{12b}$$

$$\boldsymbol{\nu}_i^{k+1} = \boldsymbol{\nu}_i^k + \alpha\left(\boldsymbol{\mu}_i^{k+1} - \boldsymbol{\zeta}^{k+1}\right), \tag{12c}$$

wherein (12a)-(12b) used the discrete version of the squared Wasserstein distance.

Replacing the squared Wasserstein distance in (8) by the entropy a.k.a. Sinkhorn regularized squared Wasserstein distance, modify the recursions (12) as

$$\begin{aligned}\boldsymbol{\mu}_i^{k+1} &= \operatorname{prox}^{W_\varepsilon}_{\frac{1}{\alpha}\left(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle\right)}\left(\boldsymbol{\zeta}^k\right)\\ &= \operatorname*{arg\,inf}_{\boldsymbol{\mu}_i \in \Delta^{N-1}} \left\{ \min_{\boldsymbol{M}\in\Pi_N(\boldsymbol{\mu}_i,\boldsymbol{\zeta}^k)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}, \boldsymbol{M}\right\rangle \right.\\ &\qquad \left. + \frac{1}{\alpha}\left(F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle\right)\right\},\end{aligned} \tag{13a}$$

$$\begin{aligned}\boldsymbol{\zeta}^{k+1} = \operatorname*{arg\,inf}_{\boldsymbol{\zeta}\in\Delta^{N-1}} \left\{ \left(\sum_{i=1}^{n} \min_{\boldsymbol{M}_i \in \Pi_N\left(\boldsymbol{\mu}_i^{k+1}, \boldsymbol{\zeta}\right)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}_i, \boldsymbol{M}_i\right\rangle \right) \right.\\ \left. - \frac{2}{\alpha}\langle \boldsymbol{\nu}_{\text{sum}}^k, \boldsymbol{\zeta}\rangle \right\},\end{aligned} \tag{13b}$$

$$\boldsymbol{\nu}_i^{k+1} = \boldsymbol{\nu}_i^k + \alpha\left(\boldsymbol{\mu}_i^{k+1} - \boldsymbol{\zeta}^{k+1}\right), \tag{13c}$$

where $\varepsilon > 0$ is a regularization parameter. In the remaining, we summarize novel results that enable us to numerically perform the recursions (13).

3. RESULTS

3.1 The $\boldsymbol{\mu}$ Update

The Sinkhorn regularized recursion (13a) allows us to get semi-analytical handle on the nested minimization via strong duality. Specifically, consider the convex functions $F_i, G_i : \Delta^{N-1} \mapsto \mathbb{R}$ for all $i \in [n]$ where $G_i(\boldsymbol{\mu_i}) := F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle$, and denote the Legendre-Fenchel conjugate of G_i as G_i^*. Following (Karlsson and Ringh, 2017, Lemma 3.5), (Caluya and Halder, 2019a, Sec. III), the Lagrange dual problem associated with (13a) is

$$\begin{aligned}\left(\boldsymbol{\lambda}_{0i}^{\text{opt}}, \boldsymbol{\lambda}_{1i}^{\text{opt}}\right) = \operatorname*{arg\,max}_{\boldsymbol{\lambda}_{0i}, \boldsymbol{\lambda}_{1i} \in \mathbb{R}^N} \left\{ \langle \boldsymbol{\lambda}_{0i}, \boldsymbol{\zeta}_k\rangle - G_i^*\left(-\boldsymbol{\lambda}_{1i}\right) \right.\\ \left. - \alpha\varepsilon\left(\exp\left(\frac{\boldsymbol{\lambda}_{0i}^\top}{\alpha\varepsilon}\right)\exp\left(-\frac{\boldsymbol{C}}{2\varepsilon}\right)\exp\left(\frac{\boldsymbol{\lambda}_{1i}}{\alpha\varepsilon}\right)\right)\right\},\ i \in [n].\end{aligned} \tag{14}$$

Using (14), the proximal updates in (13a) can then be recovered from the following proposition.

Proposition 1. ((Karlsson and Ringh, 2017, Lemma 3.5), (Caluya and Halder, 2019a, Theorem 1)) Given $\alpha, \varepsilon > 0$, the squared Euclidean distance matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$, and the probability vector $\boldsymbol{\zeta}^k \in \Delta^{N-1}$, $k \in \mathbb{N}_0$. Let $\mathbf{0}$ denote the $N \times 1$ vector of zeros. For $i \in [n]$, the vectors $\boldsymbol{\lambda}_{0i}^{\text{opt}}, \boldsymbol{\lambda}_{1i}^{\text{opt}} \in \mathbb{R}^N$ in (14) solve the system

$$\exp\left(\frac{\boldsymbol{\lambda}_{0i}^{\text{opt}}}{\alpha\varepsilon}\right) \odot \left(\exp\left(-\frac{\boldsymbol{C}}{2\varepsilon}\right)\exp\left(\frac{\boldsymbol{\lambda}_{1i}^{\text{opt}}}{\alpha\varepsilon}\right)\right) = \boldsymbol{\zeta}_k, \tag{15a}$$

$$\mathbf{0} \in \partial_{\boldsymbol{\lambda}_{1i}^{\text{opt}}} G_i^*\left(-\boldsymbol{\lambda}_{1i}^{\text{opt}}\right) - \exp\left(\frac{\boldsymbol{\lambda}_{1i}^{\text{opt}}}{\alpha\varepsilon}\right) \odot \left(\exp\left(-\frac{\boldsymbol{C}^\top}{2\varepsilon}\right)\exp\left(\frac{\boldsymbol{\lambda}_{0i}^{\text{opt}}}{\alpha\varepsilon}\right)\right). \tag{15b}$$

The proximal update $\boldsymbol{\mu}_i^{k+1}$ in (13a) is given by

$$\boldsymbol{\mu}_i^{k+1} = \exp\left(\frac{\boldsymbol{\lambda}_{1i}^{\text{opt}}}{\alpha\varepsilon}\right) \odot \left(\exp\left(-\frac{\boldsymbol{C}^\top}{2\varepsilon}\right)\exp\left(\frac{\boldsymbol{\lambda}_{0i}^{\text{opt}}}{\alpha\varepsilon}\right)\right). \tag{16}$$

We point out an important special case: if $F_i(\boldsymbol{\mu}_i) = \beta^{-1}\langle \log \boldsymbol{\mu}_i, \boldsymbol{\mu}_i\rangle$ where $\beta > 0$, then Proposition 1 reduces exactly to (Caluya and Halder, 2019a, Theorem 1) allowing further simplification of (15b). Then, (15) can be solved via certain cone-preserving block coordinate iteration proposed in (Caluya and Halder, 2019a, Sec. III.B,C) that is provably contractive. This makes the proximal update (16) semi-analytical in the sense the pair $\left(\boldsymbol{\lambda}_{0i}^{\text{opt}}, \boldsymbol{\lambda}_{1i}^{\text{opt}}\right)$ needs to be numerically computed by performing the block coordinate iteration while "freezing" the index $k \in \mathbb{N}_0$. With the converged pair $\left(\boldsymbol{\lambda}_{0i}^{\text{opt}}, \boldsymbol{\lambda}_{1i}^{\text{opt}}\right)$, the evaluation (16) is analytical for each $k \in \mathbb{N}_0$.

In our context, another case of interest is when F_i and hence G_i, is linear in $\boldsymbol{\mu}_i$. The following result shows that the proximal update $\boldsymbol{\mu}_i^{k+1}$ in this case can be computed analytically, obviating the zero order hold sub-iterations mentioned above.

Theorem 1. Given $\boldsymbol{a} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$, let $\boldsymbol{\Phi}(\boldsymbol{\mu}) := \langle \boldsymbol{a}, \boldsymbol{\mu}\rangle$ for $\boldsymbol{\mu} \in \Delta^{N-1}$. Let $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ be the squared Euclidean distance matrix, and for $\varepsilon > 0$, let $\boldsymbol{\Gamma} := \exp\left(-\boldsymbol{C}/2\varepsilon\right)$. For any $\boldsymbol{\zeta} \in \Delta^{N-1}$, $\alpha > 0$, we have

$$\operatorname{prox}^{W_\varepsilon}_{\frac{1}{\alpha}\boldsymbol{\Phi}}(\boldsymbol{\zeta}) = \exp\left(-\frac{1}{\alpha\varepsilon}\boldsymbol{a}\right) \odot \left(\boldsymbol{\Gamma}^\top\left(\boldsymbol{\zeta} \oslash \left(\boldsymbol{\Gamma}\exp\left(-\frac{1}{\alpha\varepsilon}\boldsymbol{a}\right)\right)\right)\right). \tag{17}$$

3.2 The $\boldsymbol{\zeta}$ Update

The update (13b) can be seen as a problem of computing the Sinkhorn regularized Wasserstein barycenter with an extra linear regularization. Let $W^2_{\varepsilon,\boldsymbol{\mu}_i}(\boldsymbol{\zeta}) := \min_{\boldsymbol{M}_i \in \Pi_N(\boldsymbol{\mu}_i, \boldsymbol{\zeta})} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}_i, \boldsymbol{M}_i\right\rangle$, $\varepsilon > 0$, for given $\boldsymbol{\mu}_i \in \Delta^{N-1}$ for all $i \in [n]$, and for a given squared Euclidean distance matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$. Let the superscript * denote the Legendre-Fenchel conjugate. Following (Cuturi and Peyré, 2016, Sec. 4.1), some calculations show that the dual problem corresponding to (13b) becomes

$$\begin{aligned}\left(\boldsymbol{u}_1^{\text{opt}}, \ldots, \boldsymbol{u}_n^{\text{opt}}\right) = \operatorname*{arg\,min}_{(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)\in\mathbb{R}^{nN}} \sum_{i=1}^{n} \left(W^2_{\varepsilon,\boldsymbol{\mu}_i^{k+1}}\right)^*(\boldsymbol{u}_i)\\ \text{subject to} \quad \sum_{i=1}^{n} \boldsymbol{u}_i = \frac{2}{\alpha}\boldsymbol{\nu}_{\text{sum}}^k.\end{aligned} \tag{18}$$

Consequently, the update (13b) can be performed by first solving the problem (18), and then invoking the primal-dual relation $\boldsymbol{\zeta}^{\text{opt}} = \nabla_{\boldsymbol{u}_i}\left(W^2_{\varepsilon,\boldsymbol{\mu}_i}\right)^*\left(\boldsymbol{u}_i^{\text{opt}}\right) \in \Delta^{N-1}\, \forall i \in [n]$, at the minimizer of (18). It turns out that (18) leads to an inner layer Euclidean ADMM whose structure allows efficient distributed computation.

The results summarized above lead to an overall algorithm realizing operator splitting for gradient flows in the manifold of probability measures, which solve (1) via distributed computation. Numerical experiments (not reported herein due to page constraints) on several test problems of the form (1) reveal that the proposed framework has good computational performance.

REFERENCES

Alvarez-Melis, D., Schiff, Y., and Mroueh, Y. (2021). Optimizing functionals on the space of probabilities with input convex neural networks. *arXiv preprint arXiv:2106.00774*.

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Benamou, J.D., Carlier, G., and Laborde, M. (2016). An augmented Lagrangian approach to Wasserstein gradient flows and applications. *ESAIM: Proceedings and surveys*, 54, 1–17.

Caluya, K. and Halder, A. (2021a). Wasserstein proximal algorithms for the Schrödinger bridge problem: Density control with nonlinear drift. *IEEE Transactions on Automatic Control*.

Caluya, K.F. and Halder, A. (2019a). Gradient flow algorithms for density propagation in stochastic systems. *IEEE Transactions on Automatic Control*, 65(10), 3991–4004.

Caluya, K.F. and Halder, A. (2019b). Proximal recursion for solving the Fokker-Planck equation. In *2019 American Control Conference (ACC)*, 4098–4103. IEEE.

Caluya, K.F. and Halder, A. (2021b). Reflected Schrödinger bridge: Density control with path constraints. In *2021 American Control Conference (ACC)*, 1137–1142. IEEE.

Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2), 1385–1418.

Carrillo, J.A., Craig, K., Wang, L., and Wei, C. (2021). Primal dual methods for Wasserstein gradient flows. *Foundations of Computational Mathematics*, 1–55.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31.

Chu, C., Blanchet, J., and Glynn, P. (2019). Probability functional descent: A unifying perspective on GANs, variational inference, and reinforcement learning. In *International Conference on Machine Learning*, 1213–1222. PMLR.

Cuturi, M. and Peyré, G. (2016). A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1), 320–343.

Domingo-Enrich, C., Jelassi, S., Mensch, A., Rotskoff, G., and Bruna, J. (2020). A mean-field analysis of two-player zero-sum games. *Advances in neural information processing systems*.

Halder, A. and Georgiou, T.T. (2017). Gradient flows in uncertainty propagation and filtering of linear gaussian systems. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 3081–3088. IEEE.

Halder, A. and Georgiou, T.T. (2018). Gradient flows in filtering and Fisher-Rao geometry. In *2018 Annual American Control Conference (ACC)*, 4281–4286. IEEE.

Halder, A. and Georgiou, T.T. (2019). Proximal recursion for the Wonham filter. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 660–665. IEEE.

Jarner, S.F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2), 341–361.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1), 1–17.

Karlsson, J. and Ringh, A. (2017). Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4), 1935–1962.

Mei, S., Montanari, A., and Nguyen, P.M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33), E7665–E7671.

Mokrov, P., Korotin, A., Li, L., Genevay, A., Solomon, J., and Burnaev, E. (2021). Large-scale Wasserstein gradient flows. *arXiv preprint arXiv:2106.00736*.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in optimization*, 1(3), 127–239.

Peyré, G. (2015). Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4), 2323–2351.

Roberts, G.O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4), 337–357.

Rotskoff, G.M. and Vanden-Eijnden, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050, 22.

Sirignano, J. and Spiliopoulos, K. (2020). Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3), 1820–1852.

Stramer, O. and Tweedie, R. (1999a). Langevin-type models i: Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, 1(3), 283–306.

Stramer, O. and Tweedie, R. (1999b). Langevin-type models ii: Self-targeting candidates for MCMC algorithms. *Methodology and Computing in Applied Probability*, 1(3), 307–328.

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Soc., 1st edition.

Zhang, J., Koppel, A., Bedi, A.S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33, 4572–4583.

Zhang, R., Chen, C., Li, C., and Carin, L. (2018). Policy optimization as Wasserstein gradient flows. In *International Conference on Machine Learning*, 5737–5746. PMLR.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On Scale Fragilities in Localized Consensus

Emma Tegling *

* *Department of Automatic Control, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden*(emma.tegling@control.lth.se).

Abstract: We consider the prototypical networked control problem of distributed consensus in networks of agents with integrator dynamics of order two or higher ($n \geq 2$). We assume all feedback to be localized in the sense that each agent has a bounded number of neighbors and consider a scaling of the network through the addition of agents. We show that standard consensus algorithms that rely on relative state feedback and fixed gains can be subject to scale fragilities, meaning that stability is lost as the network grows. For high-order agents ($n \geq 3$), we prove that no consensus algorithm is what we term *scalably stable*. That is, while a given algorithm may allow a small network to converge, it causes instability if the network grows beyond a certain finite size. This holds in families of network graphs whose algebraic connectivity, that is, the smallest non-zero Laplacian eigenvalue, is decreasing towards zero in network size (equivalently, non-expanding graphs). For second-order consensus ($n = 2$), we prove that the same scale fragility applies to classes of directed graphs that have a complex Laplacian eigenvalue approaching the origin (e.g. directed ring graphs). We derive algebraic conditions for the affected graphs, and discuss how the consensus algorithm can be modified to retrieve scalable stability.

Keywords: Distributed control, large-scale systems, robustness, algebraic graph theory. AMS Subject classification: 93A14, 93A15.

1. INTRODUCTION

Characterizing the dynamic behaviors of networked or multi-agent systems has been an active research area for many years. In particular, since the works by Fax and Murray (2004), Olfati-Saber and Murray (2004), and Jadbabaie et al. (2003), the prototypical sub-problem of distributed consensus has been the subject of significant research efforts. While the particular modeling aspects vary, the consensus objective is to coordinate agents in a network to a common state of agreement. Engineering applications range from distributed computing and sensing to power grid synchronization and coordination of vehicles.

In this work, we consider a consensus algorithm of order n, where each agent $i \in \{1, 2, \dots, N\}$ in a network is modeled as an n^{th} order integrator, and the control input is a weighted sum of relative feedback terms with respect to the agent's neighbors. That is,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} x_i^{(0)}(t) &= x_i^{(1)}(t) \\ &\vdots \\ \frac{\mathrm{d}}{\mathrm{d}t} x_i^{(n-2)}(t) &= x_i^{(n-1)}(t) \\ \frac{\mathrm{d}}{\mathrm{d}t} x_i^{(n-1)}(t) &= u_i(t), \end{aligned} \tag{1}$$

with the state $x_i^{(0)}(t) = x_i(t) \in \mathbb{R}$, and

$$u_i = -\sum_{k=0}^{n-1} a_k \sum_{j \in \mathcal{N}_i} w_{ij}(x_i^{(k)} - x_j^{(k)}), \tag{2}$$

where the $a_k > 0$ are fixed gains and w_{ij} are edge weights in the network graph. For $n = 1$ this reduces to the familiar first-order, or information, consensus algorithm. For $n = 2$, we obtain second-order consensus, which is often used to model formation control in multi-vehicle networks. The problem for $n \geq 3$, to which several results in this work pertain, has also received significant attention, see e.g. Ren et al. (2007); Rezaee and Abdollahi (2015); Zuo et al. (2018). This can be viewed as an important theoretical generalization of the first- and second-order algorithms, but also has practical relevance. For example, position, velocity, as well as acceleration feedback play a role in flocking behaviors, resulting in a model where $n = 3$ Ren et al. (2007).

Existing literature has typically focused on deriving conditions for convergence of a given set of agents to consensus, and how such conditions depend on various properties of the network. We take a different perspective and focus on the *scalability* of given consensus algorithm (2) to ever larger networks. In other words, we assume that interactions between agents are *fixed* (i.e., pre-designed) and *localized*, and grow the network through the addition of more and more agents. Formally, we model the consensus algorithm over a *family* of network graphs $\{\mathcal{G}_N\}_{N\to\infty}$ with a common upper bound on nodal degrees.

We show that consensus of order $n \geq 2$ is subject to a *scale fragility* in certain graph families. This implies that stability (and thereby convergence to consensus) is lost if

⋆ This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and the Swedish Research Council through grant 2019-00691.

the network grows beyond some *finite* size. For $n \geq 3$, our result is particularly clear-cut: the consensus algorithm (2) lacks what we term *scalable stability* in any family of graphs whose algebraic connectivity decreases towards zero in network size N. This is true in any *non-expanding* graph family, meaning all bounded-degree graphs where connections are, in a sense, localized.

For second-order consensus ($n = 2$), the scale fragility applies to particular classes of directed graphs with a complex Laplacian eigenvalue that approaches the origin as N increases. This includes, for example, directed ring graphs. The particular result for ring graphs has previously been reported in Cantos et al. (2016); Herman (2016), but our work provides a generalization. The result implies that ring-shaped vehicular formations that, e.g, use adaptive cruise control modeled as in Gunter et al. (2021), are at risk of becoming unstable.

The key results summarized in this extended abstract are presented in detail in Tegling et al. (2022) with preliminary versions appearing in Tegling et al. (2019a,b) .

2. NOTATION AND PROBLEM SETUP

Consider a network modeled by the graph $\mathcal{G}_N = \{\mathcal{V}_N, \mathcal{E}_N\}$ with $N = |\mathcal{V}_N|$ nodes. The set $\mathcal{E}_N \subset \mathcal{V}_N \times \mathcal{V}_N$ contains the edges, each of which has an associated nonnegative weight w_{ij}. The graph $\mathcal{G}_N$, which in general will be a directed graph, is a member of a sequence, or family, of graphs $\{\mathcal{G}_N\}_{N\to\infty}$. We remark that $\mathcal{G}_N$ need not be a subgraph of $\mathcal{G}_{N+1}$. The graph Laplacian L of $\mathcal{G}_N$ is defined as follows:

$$[L]_{ij} = \begin{cases} -w_{ij} & \text{if } j \neq i \text{ and } j \in \mathcal{N}_i \\ \sum_{k\in\mathcal{N}_i} w_{ik} & \text{if } j = i \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathcal{N}_i$ defines the neighborhood of node $i \in \mathcal{V}_N$, that is, the set of nodes j such that $(i, j) \in \mathcal{E}_N$. Denote by λ_l (or $\lambda_l(\mathcal{G}_N)$ where explicitness is needed) with $l = 1, \ldots, N$ the eigenvalues of L. Zero is a simple eigenvalue of L if and only if the graph has a connected spanning tree, which we assume henceforth. Remaining eigenvalues are in the complex right half plane (RHP), and numbered so that $0 = \lambda_1 < \mathrm{Re}\{\lambda_2\} \leq \ldots \leq \mathrm{Re}\{\lambda_N\}$.

Defining the system state $\xi = [x^{(0)}, x^{(1)}, \ldots, x^{(n-1)}]^T \in \mathbb{R}^{Nn}$ and making use of L, we can write the system's closed-loop dynamics (1)–(2) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\xi = \underbrace{\begin{bmatrix} 0 & I_N & 0 & \cdots & 0 \\ 0 & 0 & I_N & \cdots & \vdots \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I_N \\ -a_0 L & -a_1 L & -a_2 L & \cdots & -a_{\mathrm{n}-1} L \end{bmatrix}}_{\mathcal{A}} \xi. \tag{4}$$

3. RESULTS

We first state a number of important assumptions that underlie our analysis:

Assumption 1. (Finite gains). The controller gains are finite, that is, $a_k \leq a_{\max} < \infty$ for all $k = 0, 1, \ldots, n$.

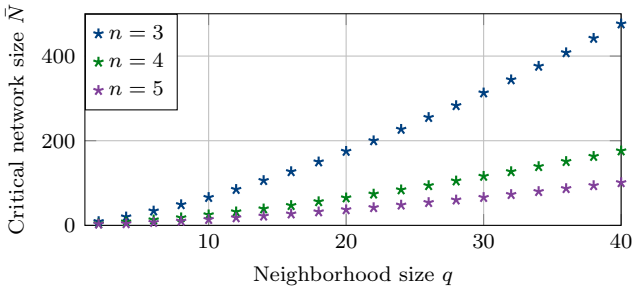


Fig. 1. Critical network size $\bar{N}$ for an n^{th} order consensus algorithm. The graph is an undirected path graph where each node is connected to its q nearest neighbors. Increasing the neighborhood size q here increases $\bar{N}$ faster than linearly (we also derive the exact scaling).

Assumption 2. (Fixed gains). The gains a_k for all $k = 0, 1, \ldots, n$ do not change if the underlying graph changes. That is, the gains are fixed with respect to the graph family $\{\mathcal{G}_N\}_{N\to\infty}$. In particular, they are independent of N.

We will impose the following assumptions on the network graph:

Assumption 3. (Bounded neighborhoods). All nodes in the graph family $\{\mathcal{G}_N\}_{N\to\infty}$ have a neighborhood of size at most q, where q is fixed and independent of N. That is,

$$|\mathcal{N}_i| \leq q \quad \forall i \in \mathcal{V}_N. \tag{5}$$

Assumption 4. (Finite weights). The edge weights in each $\mathcal{G}_N$ are finite, that is, $w_{ij} \leq w_{\max} < \infty$ for all $(i, j) \in \mathcal{E}_N$, where $w_{\max}$ is fixed and independent of N.

Assumptions 3–4 imply that we consider networks with bounded nodal degrees.

3.1 Scalable stability in high-order consensus

The network of agents is said to be achieving consensus if $x_i^{(k)} \to x_j^{(k)}$ for all $i, j \in \mathcal{V}_N$, all $k = 0, 1, \ldots, n-1$, and for any initial state. It is known that the algorithm (2) achieves consensus if the eigenvalues of the system matrix $\mathcal{A}$ defined in (4) are in the left half plane, apart from exactly n zero eigenvalues that are associated with the drift of the network average Ren et al. (2007). We focus on a scenario where these conditions may hold for small network sizes N, but where one or more eigenvalues leaves the left half plane and causes instability as the network grows beyond some network size $\bar{N}$. In these cases, we say the control algorithm lacks *scalable stability*.

Definition 1. (Scalable stability). A consensus control design is *scalably stable* if the resulting closed-loop system achieves consensus over *any* graph in the family $\{\mathcal{G}_N\}_{N\to\infty}$ of finite size N.

We are now ready to summarize our main results. First, we prove that high order ($n \geq 3$) consensus is subject to a scale fragility.

Theorem 3.1. If $n \geq 3$, no control on the form (2) subject to Assumptions 1–2, is scalably stable in graph families where the sequence $\mathrm{Re}\{\lambda_2(\mathcal{G}_N)\} \to 0$ as $N \to \infty$.

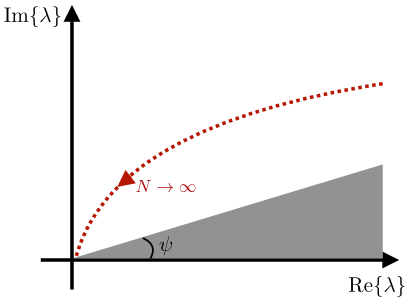


Fig. 2. Illustration of the condition in Theorem 3.2. The example trajectory shows λ_2 of a directed ring graph.

The proof relies on the Routh-Hurwitz criteria for polynomials with complex-valued coefficients. It is found in Tegling et al. (2022).

Theorem 3.1 implies that high-order consensus does not scale in graph families where the algebraic connectivity is decreasing in network size. This applies to all bounded-degree graph families except expander graphs (see Section 3.2). Instability will occur at the smallest size N for which the Routh-Hurwitz criteria are violated, and at least one eigenvalue crosses to the RHP. We will denote this critical network size $\bar{N}$. In Figure 1 we display $\bar{N}$ for $n = 3, 4, 5$ in an unweighted path graph.

Remark 1. In high-order leader-follower consensus, scalable stability is never attained in undirected bounded-degree graph families. This is because the smallest eigenvalue of the *grounded* graph Laplacian $\bar{\lambda}_1(\mathcal{G}_N) \leq \frac{q}{N-1} w_{\max} \to 0$. This also implies that consensus over any expander graph family (where λ_2 remains large, despite bounded nodal degrees) is *fragile* towards grounding, i.e., towards one agent becoming a leader, since $\bar{\lambda}_1(\mathcal{G}_N) \ll \lambda_2(\mathcal{G}_N)$.

Second-order consensus lacks scalable stability in certain families of *directed* graphs with complex eigenvalues:

Theorem 3.2. If $n \geq 2$, no control on the form (2), subject to Assumptions 1–2, is scalably stable in graph families where, for a fixed index $\bar{l} < N$,

(1) $\mathrm{Re}\{\lambda_{\bar{l}}(\mathcal{G}_N)\} \to 0$ as $N \to \infty$, *and*
(2) for each N and at least one $l \in \{2, 3, ..., \bar{l}\}$ it holds $\arg\{\lambda_l(\mathcal{G}_N)\} > \psi$, where $\psi \in (0, \pi/2)$ is a constant angle independent of N.

A particular graph family where Theorem 3.2 applies is directed ring graphs (a ring graph that is not undirected) with uniform edge weights. This was already observed in Cantos et al. (2016); Herman (2016). We show that it extends to toric lattices and their fuzzes. More generally, however, it is an open graph-theoretical challenge [1] to characterize graph families that have complex-valued eigenvalues and can be affected by the scale fragility we describe.

3.2 An algebraic condition for non-expanding graphs

In Theorem 3.1, we showed that high-order consensus lacks stable scalability in any graph family that has an algebraic connectivity that decreases towards zero as the network grows. This is true in any undirected graph family that is not an expander family. Expander families are defined by an isoperimetric constant that is lower bounded. We do not include the formal definition, which is combinatoric, here, but refer the reader to, e.g., Chung (1997). Intuitively, a large isoperimetric constant implies that every part of the graph is well interconnected with every other – the graph has no "bottleneck". Instead, we propose a novel algebraic condition for undirected graphs that have a bottleneck and are therefore not expander graphs.

For this purpose, partition a graph's vertex set into three disjoint sets X_0, X_1, X_2 so that $X_0 \cup X_1 \cup X_2 = \mathcal{V}$ and $|X_0| = N_0$,$|X_1| = N_1$, $|X_2| = N_2$. Each node in X_0 is connected to at least one node in both X_1 and X_2, but no edges connect X_1 and X_2 directly. See also Fig. 3. In other words, X_0 is the boundary set of both X_1 and X_2. This partitioning is always possible, unless the graph is complete (note that X_0, X_1, X_2 need not be connected subgraphs).

By re-numbering the nodes, the graph Laplacian becomes

$$L = \begin{bmatrix} L_1 & L_{10} & 0_{N_1 \times N_2} \\ L_{10}^T & L_0 & L_{03}^T \\ 0_{N_2 \times N_1} & L_{30} & L_2 \end{bmatrix}. \tag{6}$$

If N_0 can be made small in relation to both N_1 and N_2, we say that the graph has a bottleneck. The following lemma shows that if the bottleneck remains as the network grows, then $\{\lambda_2(\mathcal{G}_N)\} \to 0$.

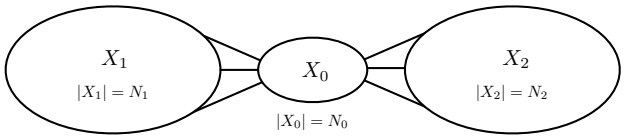


Fig. 3. Illustration for Lemma 3.3.

Lemma 3.3. Let $\{\mathcal{G}_N\}_{N\to\infty}$ be an undirected graph family subject to Assumptions 3–4. If every graph $\mathcal{G}_N$ in the family can be partitioned as outlined above in such a way that $N_0/N_1 \to 0$ and $N_0/N_2 \to 0$ as $N \to \infty$, then $\{\lambda_2(\mathcal{G}_N)\} \to 0$ as $N \to \infty$.

The proof relies on the Rayleigh-Ritz theorem and the assumption on bounded nodal degrees.

Remark 2. The specific scaling of $\lambda_2(\mathcal{G}_N)$ in N is known for several classes of graphs. For example, for planar graphs, $\lambda_2(\mathcal{G}_N) \leq \frac{8qw_{\max}}{N}$.

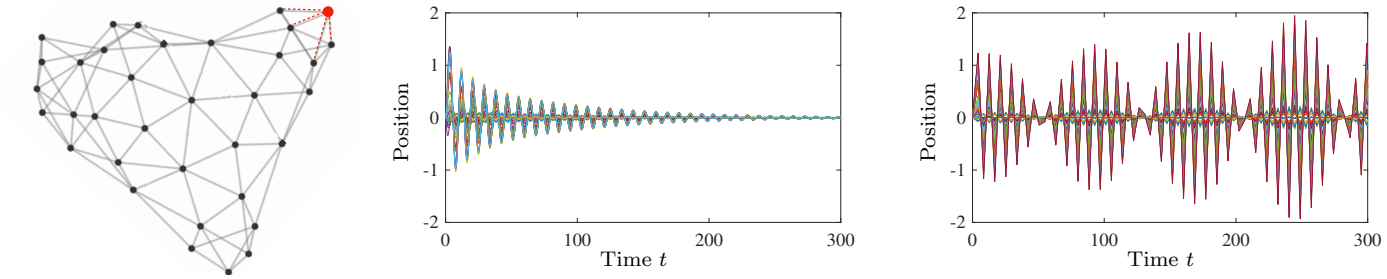
Example 1. To illustrate Theorem 3.1, we consider a third-order consensus algorithm:

$$x_i^{(3)} = -\sum_{j\in\mathcal{N}_i} \left[0.5(x_i - x_j) + (\dot{x}_i - \dot{x}_j) + (\ddot{x}_i - \ddot{x}_j)\right],$$

over the 34-node graph depicted in Figure 4a with unit edge weights. Here, $\lambda_2(\mathcal{G}_{34}) = 0.536$ and the system achieves consensus. Adding a 35^{th} node along with 4 connecting edges gives $\lambda_2(\mathcal{G}_{35}) = 0.493$ and the system becomes unstable, see Fig. 4c. [2]

[1] One we hope to discuss with experts at MTNS 2022!

[2] This particular value for $\lambda_2(\mathcal{G}_{35})$ depends on the placement of the 35^{th} node. Other placements can allow the critial $\bar{N} > 35$, but instability occurs eventually.



(a) Network graph (b) $N = 34$ (c) $N = 35$

Fig. 4. Simulation of 3rd order consensus over graph depicted in (a) subject to random initial accelerations. In (b) the network's 34 agents converge to an equilibrium. In (c) a 35th node has been added, indicated by red color in the graph. This addition leads to instability. The plots (b) and (c) show position trajectories relative to Agent no. 1.

4. DISCUSSION

The scale fragilities we describe here can in principle be attributed to two model features. First, the relative state feedback upon which the consensus algorithm is based. It is known that a restriction to relative feedback imposes performance and design limitations; an issue that was recently analyzed formally in Jensen and Bamieh (2022). In our work, we discuss how scalability can be retrieved if the controller has access to absolute feedback.

Second, the locality property, that is, bounded nodal neighborhoods, is key for our results. A natural question is therefore how nodal neighborhoods would need to scale to alleviate the scale fragility. Interestingly, we can prove that it can suffice to grow neighborhoods as $q \sim N^{2/3}$. We note that this only holds for leaderless consensus; leader-follower consensus still requires neighborhoods proportional to N.

We remark that, in order to be able to discuss a given controller's scalability in a network of increasing size, the assumption that it be fixed is necessary. That is, the controller cannot be re-tuned as the network grows. By re-tuning the consensus algorithm (2), either by changing the gains a_k, weights w_{ij}, or by relaxing the locality assumption, scalable stability can be retrieved. Such a re-tuning requires knowledge of *global* properties of the system. Still, the design of controller re-tuning protocols is an interesting direction for future research.

ACKNOWLEDGEMENTS

Collaboration and helpful discussions with Henrik Sandberg, Bassam Bamieh, Maria Seron, and Rick Middleton are gratefully acknowledged.

REFERENCES

Cantos, C., Veerman, J., and Hammond, D. (2016). Signal velocity in oscillator arrays. *Eur. Phys. J. Spec.*, 225(6), 1115–1126.

Chung, F. (1997). *Spectral Graph Theory*. Providence, RI.

Fax, J.A. and Murray, R.M. (2004). Information flow and cooperative control of vehicle formations. *IEEE Trans. Autom. Control*, 49(9), 1465–1476.

Gunter, G., Gloudemans, D., Stern, R.E., McQuade, S., Bhadani, R., Bunting, M., Delle Monache, M.L., Lysecky, R., Seibold, B., Sprinkle, J., Piccoli, B., and Work, D.B. (2021). Are commercially implemented adaptive cruise control systems string stable? *IEEE Trans. Intell. Transp. Syst.*, 22(11), 6992–7003.

Herman, I. (2016). *Scaling in vehicle platoons*. Phd thesis, Czech Technical University in Prague. URL `https://support.dce.felk.cvut.cz/mediawiki/images/d/d1/Diz\_2017\_herman\_ivo.pdf`.

Jadbabaie, A., Lin, J., and Morse, A.S. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Trans. Autom. Control*, 48(6), 988–1001.

Jensen, E. and Bamieh, B. (2022). On structured-closed-loop versus structured-controller design: the case of relative measurement feedback.

Olfati-Saber, R. and Murray, R.M. (2004). Consensus problems in networks of agents with switching topology and time-delays. *IEEE Trans. Autom. Control*, 49(9), 1520–1533.

Ren, W., Moore, K.L., and Chen, Y. (2007). High-order and model reference consensus algorithms in cooperative control of multi-vehicle systems. *J. Dyn. Syst. Meas. Control*, 129(5), 678–688.

Rezaee, H. and Abdollahi, F. (2015). Average consensus over high-order multiagent systems. *IEEE Trans. Autom. Control*, 60(11), 3047–3052.

Tegling, E., Bamieh, B., and Sandberg, H. (2022). Scale fragilities in localized consensus dynamics. *arXiv preprint arXiv:2203.11708*.

Tegling, E., Bamieh, B., and Sandberg, H. (2019a). Localized high-order consensus destabilizes large-scale networks. In *American Control Conf. (ACC)*, 760–765.

Tegling, E., Middleton, R.H., and Seron, M.M. (2019b). Scalability and fragility in bounded-degree consensus networks. In *8th IFAC Workshop on Distributed Estimation and Control in Networked Systems (NecSys)*.

Zuo, Z., Tian, B., Defoort, M., and Ding, Z. (2018). Fixed-time consensus tracking for multi-agent systems with high-order integrator dynamics. *IEEE Trans. Autom. Control*, 63(2), 563–570.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Optimal control of parabolic equations – a spectral calculus based approach ⋆

Martin Lazar * **Luka Grubišić** ** **Ivica Nakić** **
Martin Tautenhahn ***

* *University of Dubrovnik, Department of Electrical Engineering and Computing, Croatia*
** *University of Zagreb, Department of Mathematics, Croatia*
*** *Universität Leipzig, Fakultät für Mathematik und Informatik, Germany*

Abstract: Recent theoretical and numerical results on a constrained optimal control problem of Bolza type for a class of parabolic equations obtained by the authors are presented. We study the case when the cost functional is quadratic and comprises the norm of a control and the distance of the system trajectory from the desired evolution profile, the constraint is imposed on the final state that should be steered within a prescribed distance to a given target and the control enters the system through the initial condition.The theoretical results provide a formula for the optimal control and the numerical algorithm is based on efficient rational Krylov approximation techniques.

Keywords: optimal control; parabolic equations; homogenization; rational Krylov spaces; spectral calculus

1. INTRODUCTION

We consider a sequence of optimal control problems associated to the heat equation with rapidly oscillating coefficients. The cost functional to be minimised is quadratic and comprises the norm of a control and the distance of the system trajectory from the desired evolution profile. The constraint is imposed on the final state at time $T > 0$ that should be steered within a prescribed distance to a given target.

The control enters the system through the initial condition, and we deal with an inverse problem (of initial source identification) from the optimal control viewpoint. Initial control problems for parabolic equations are less investigated than distributed or boundary ones. A reason for that is due to the strong dissipativity of parabolic equations, which make the inverse problem numerically very challenging. The problem has been tackled by different approaches Fabre et al. (1995); Meidner and Vexler (2007); Li et al. (2014); Casas et al. (2015). In order to successfully address these difficulties we apply a recently developed approach, based on the spectral calculus for self-adjoint operators and a geometrical representation of the problem (Grubišić et al. (2021)). In the paper the authors first obtained a closed-form expression for the control solution as a function of the self-adjoint operator governing the dynamics of the system. The numerical computations were achieved by exploring efficient rational Krylov approximation techniques for resolvents from Berljafa and Güttel (2017), by which one constructs a rational approximant of the aforementioned function of the operator.

⋆ The research was done while the first author was visiting Chair of Dynamics, Control and Numerics (Alexander von Humboldt Professorship) at Friedrich-Alexander-Universität Erlangen-Nürnberg, with the support of the DAAD (Research Stays for University Academics and Scientists, 2021 programme) and Alexander von Humboldt-Professorship.

2. PROBLEM FORMULATION AND THE SOLUTION FORMULA

Let A be a self–adjoint operator bounded below on an infinite dimensional Hilbert space $\mathcal{H}$. For $u \in \mathcal{H}$ we consider the Cauchy problem

$$y'(t) + Ay(t) = 0, \quad t > 0,$$
$$y(0) = u.$$

By $(S_t)_{t\geq 0}$ we denote the semigroup generated by $-A$.

The optimal control problem Given $\epsilon, T > 0$ and $y^* \in \mathcal{H}$ we introduce the constrained minimisation problem

$$\min_{u\in\mathcal{H}} \left\{ J(u) \colon \left\| S_T u - y^* \right\| \leq \epsilon \right\} \tag{1}$$

where

$$J(u) = \frac{\alpha}{2}\|u\|^2 + \frac{1}{2}\int_0^T \beta(t) \left\| S_t u - w(t) \right\|^2 \mathrm{d}t, \tag{2}$$

$\alpha > 0$ and $\beta \in L^\infty((0,T);[0,\infty))$ are weights of the cost, and $w \in L^2((0,T);\mathcal{H})$ is the target trajectory. □

By using classical convex optimization techniques (e.g. (Peypouquet, 2015, Section 3.6)) one can show that the problem is well posed and admits the unique solution that we denote by $\hat{u}$.

If the solution of the unconstrained problem

$$\tilde{u} = \min_{u\in\mathcal{H}} \{J(u)\}$$

drives the system to the ε ball around the target, then the solutions of the two problems coincide and the original

problem can be relaxed by considering the unconstrained one, which is easier to handle. This case we exclude from further analysis.

If the problem can not be relaxed to the corresponding unconstrained one, then the optimal final state lies on the boundary of the target ball (Lazar et al., 2017, Proposition 2.1). In such a way one can associate a Lagrange functional to the problem and obtain the formula for the solution. The formula was first obtained in Lazar et al. (2017) and subsequently further developed and elaborated in details in Grubišić et al. (2021).

Theorem 1. The optimal initial state is given by

$$\hat{u} = (\mu^\epsilon S_{2T} + \Psi)^{-1}(\mu^\epsilon S_T y^* + \psi), \tag{3}$$

where

$$\Psi = \alpha\,\mathrm{Id} + \int_0^T \beta(t) S_{2t} dt, \quad \psi = \int_0^T \beta(t) S_t w(t) dt,$$

and $\mu^\epsilon \geq 0$ is the unique solution of

$$\Phi(\mu) = \epsilon \tag{4}$$

if $\epsilon < \|\tilde{y} - y^*\| = \|\Psi^{-1} S_T \psi - y^*\|$, and zero otherwise. Here $\Phi\colon [0,\infty) \to [0,\infty)$ is the function defined by

$$\Phi(\mu) = \|y^* - (\mu S_{2T} + \Psi)^{-1}(\mu S_{2T} y^* + S_T \psi)\|. \tag{5}$$

□

Formula (3) is almost explicit, up to the scalar μ^ϵ which corresponds to the optimal Lagrange multiplier.

3. NUMERICAL IMPLEMENTATION OF THE SOLUTION OF THE OPTIMAL CONTROL PROBLEM

In the numerical implementation we first devise a method for solving the equation (4). Hereby we explore efficient Krylov subspace techniques that allow us to approximate a (generalised) exponential functions of an operator appearing in the equation (5) by a series of linear problems.

The approximation is obtained by using the award winning `rkfit` algorithm from Berljafa and Güttel (2017). This is a rational Krylov function fitting algorithm which has also been implemented in Matlab within the Rational Krylov Toolbox.

The algorithm approximates a function of an operator by a rational function r, hereby using the equality

$$(v, S_t v) - (v, r(A)v) = \int_{-\infty}^{\kappa} \left(\mathrm{e}^{t\lambda} - r(\lambda)\right) \mathrm{d}(E(\lambda)v, v), \tag{6}$$

where $E(\cdot)$ denotes the spectral measure of the self-adjoint operator A (Kato (1995)).

In the next step the rational function r is rewritten in a partial fractions form

$$r(z) = r_0 + \frac{r_1}{z - \zeta_1} + \cdots \frac{r_d}{z - \zeta_d}.$$

This provides application of a function of the operator A of the form

$$S_t \approx r_0 I + \sum_{i=1}^{d} r_i (A - \zeta_i)^{-1},$$

where $\zeta_i,\ i = 1, \ldots, d$, belong to the resolvent set of A.

Once the above approximations are provided, one can use any root finding algorithm based only on function evaluation to robustly approximate the root of (4).

The optimal control is then obtained by solving the linear equation (3). Note that the expressions entering it are of the same form as those appearing in the equation for the optimal Lagrange multiplier (4).

The efficiency of the procedure was confirmed by numerical examples, including several constrained optimization problems in 1D and 2D, with variable and non-smooth coefficients and acting on irregular domains. ALso the sensitivity analysis of the solution was provided.

REFERENCES

Berljafa, M. and Güttel, S. (2017). The RKFIT algorithm for nonlinear rational approximation. *SIAM J. Sci. Comput.*, 39(5), A2049–A2071. doi: 10.1137/15M1025426.

Casas, E., Vexler, B., and Zuazua, E. (2015). Sparse initial data identification for parabolic PDE and its finite element approximations. *Math. Control Relat. Fields*, 5(3), 377–399. doi:10.3934/mcrf.2015.5.377.

Fabre, C., Puel, J., and Zuazua, E. (1995). On the density of the range of the semigroup for semilinear heat equations. In J. Lagnese, D. Russell, and L. White (eds.), *Control and Optimal Design of Distributed Parameter Systems*, volume 70 of *The IMA Volumes in Mathematics and its Applications*, 73–91. Springer, New York.

Grubišić, L., Lazar, M., Nakić, I., and Tautenhahn, M. (2021). Optimal control of parabolic equations – a spectral calculus based approach. *submitted*, 25 pp.

Kato, T. (1995). *Perturbation theory for linear operators*, volume 132 of *Classics in Mathematics*. Springer, Berlin. doi:10.1007/978-3-642-66282-9.

Lazar, M., Molinari, C., and Peypouquet, J. (2017). Optimal control of parabolic equations by spectral decomposition. *Optimization*, 66(8), 1359–1381.

Lazar, M. and Zuazua, E. (2022). Greedy search of optimal approximate solutions. *submitted*, 17 pp.

Li, Y., Osher, S., and Tsai, R. (2014). Heat source identification based on l_1 constrained minimization. *Inverse Probl. Imaging*, 8(1), 199–221. doi: 10.3934/ipi.2014.8.199.

Meidner, D. and Vexler, B. (2007). Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.*, 46(1), 116–142. doi: 10.1137/060648994.

Peypouquet, J. (2015). *Convex optimization in normed spaces: theory, methods and examples*. SpringerBriefs in Optimization. Springer, Cham. doi:10.1007/978-3-319-13710-0.

Zuazua, E. (1994). Approximate controllability for linear parabolic equations with rapidly oscillating coefficients. *Control Cybernet.*, 23(4), 793–801. Modelling, identification, sensitivity analysis and control of structures.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Port-Hamiltonian modeling of power networks with distributed transmission lines

Hannes Gernandt * **Dorothea Hinsen** *

* *Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin (e-mail: {gernandt, hinsen}@math.tu-berlin.de).*

Abstract: In this talk we consider power networks consisting of loads and generators which are interconnected via transmission lines. Here we use a distributed model for the transmission lines and provide a port-Hamiltonian formulation as a boundary control system and show the exponential stability and a power-balance equation for classical solutions.

Keywords: Distributed port-Hamiltonian systems, Infinite Dimensional Systems Theory, Systems on Graphs, Descriptor Systems, Power Grids

1. OUTLINE

We consider power grid models consisting of loads and generators which are interconnected by transmission lines. Here we follow the port-Hamiltonian (pH) modeling approach to power grids which was developed in Fiaz et al. (2012); Fiaz et al. (2013); van der Schaft and Stegink (2016).

As a new aspect we consider distributed models for the transmission lines based on the telegrapher's equations. Although this model has been studied in Jacob and Zwart (2012a), to the best of our knowledge no interconnection of several transmission lines has been studied yet. Hence as a first contribution, we formulate the interconnected system of transmission lines and loads as a boundary control system and study the exponential stability of the semigroups. Second, we study the corresponding finite-dimensional port-Hamiltonian systems which are obtained after a spatial discretization of the transmission lines and add further port-Hamiltonian power generator models

2. PORT-HAMILTONIAN OPERATOR CLASS

The distributed transmission line on the spatial interval $[0,\ell]$ can be modeled by the telegraphers equations

$$\begin{aligned} C(\xi)\tfrac{\partial}{\partial t}V(t,\xi) &= -\tfrac{\partial}{\partial \xi}I(t,\xi) - G(\xi)V(t,\xi),\\ L(\xi)\tfrac{\partial}{\partial t}I(t,\xi) &= -\tfrac{\partial}{\partial \xi}V(t,\xi) - R(\xi)I(t,\xi), \qquad (1)\\ V(0,\xi) &= V^0(\xi), \quad I(0,\xi) = I^0(\xi) \end{aligned}$$

where $I : [0,\ell] \to \mathbb{R}^d$ is the current and $V : [0,\ell] \to \mathbb{R}^d$ is the voltage across the transmission line. We allow values in $\mathbb{R}^d$ for some $d \geq 1$ to include three phase models of transmission lines as well. Furthermore, $C(\xi) \geq c_0$ and $L(\xi) \geq l_0$ for all $\xi \in [0,\ell]$ and some $c_0, l_0 > 0$, $R(\xi) \geq 0$ and $G(\xi) \geq 0$.

This motivates the following modifications and generalizations of the pH model:

(M1) We consider an arbitrary Hilbert space X with scalar product $\langle \cdot, \cdot \rangle$ and define on $X \times X$ the block operator

$$A = \begin{bmatrix} -G & D \\ D & -R \end{bmatrix}$$

where $D : X \supset \operatorname{dom} D \to X$ is closed, densely defined and skew-symmetric, i.e. $D \subseteq -D^*$. Furthermore, we assume that G, R are bounded and fulfill $\langle Rx, x\rangle \geq 0$ and $\langle Gx, x\rangle \geq 0$ for all $x \in X$;

(M2) $\mathcal{H} : X \times X \to X \times X$ is bounded and fulfills $\langle \mathcal{H}x, x\rangle \geq m\|x\|^2$;

(M3) Instead of specifying the domain of A in terms of the boundary flow and boundary effort f_∂ and e_∂, we assume that the symmetric operator $\mathrm{i}D$ has a so called *boundary triplet* $\{\mathcal{X}, \Gamma_0, \Gamma_1\}$, see e.g. Behrndt et al. (2020).

Assuming (M1)-(M3), we consider the following boundary control system

$$\dot{x}(t) = \begin{bmatrix} -G & D^* \\ D^* & -R \end{bmatrix} \mathcal{H}x, \quad \Gamma_0\mathcal{H}x(t) = u(t), \quad \Gamma_1\mathcal{H}x(t) = y(t) \qquad (2)$$

where the mappings $\Gamma_0, \Gamma_1 : \operatorname{dom}(A^*\mathcal{H}) \to \mathcal{X}$ are given by a boundary triplet. Such systems were studied in a more general context in Malinen and Staffans (2006, 2007) and we demonstrate how their results can be applied to derive a power-balance equation for classical solutions x of (2)

$$\begin{aligned} &\langle \mathcal{H}x(t), x(t)\rangle - \langle \mathcal{H}x(0), x(0)\rangle \\ &\leq 2\int_0^t \operatorname{Re}\langle u(\tau), y(\tau)\rangle d\tau - \left\langle \begin{bmatrix} R & 0 \\ 0 & G \end{bmatrix} \mathcal{H}x(t), \mathcal{H}x(t) \right\rangle. \qquad (3) \end{aligned}$$

The power balance equation (3) implies that the considered boundary control systems are *impedance passive system* in the sense of Staffans (2002). For general systems which can be decomposed into a skew-adjoint and a dissipating part, and with bounded control operators a power balance equation for mild solutions was given in Philipp et al. (2021), see also Egger et al. (2018) for power balance equations for weak solutions.

Furthermore, we show that the operator

$$A := \begin{bmatrix} -G & D^* \\ D^* & -R \end{bmatrix}|_{\ker(\Gamma_0\mathcal{H})} = \begin{bmatrix} -G & -D^* \\ D & -R \end{bmatrix}$$

generates an exponentially stable semigroup if R is *uniformly positive*, i.e. there exists $r_0 > 0$ such that

$$\langle Rx, x\rangle \geq r_0\|x\|^2, \quad \text{for all } x \in X$$

together with either the surjectivity of D^*, or compact resolvent assumptions which are typically fulfilled in many transport network examples or beam networks. Note that if G is uniformly positive as well, then the exponential stability trivially follows from the well-known Lyapunov inequality.

The exponential stability and stabilizability for pH systems is well studied, see e.g. Villegas (2007); Villegas et al. (2009); Augner and Jacob (2014); Ramirez et al. (2014); Augner (2020); Trostorff and Waurick (2022) and in particular (Jacob and Zwart, 2012b, Chapter 9). But in these works the characterizations are mostly in terms of conditions on the boundary values of the Hamiltonian. Furthermore, there are some recent results Skrepek (2021) on stability of multi-dimensional pH systems. In particular, there are results on the strong and semi-uniform stability, see also Kurula and Zwart (2015) for a treatment of multi-dimensional pH systems.

3. NETWORK OF TRANSMISSION LINES

We consider a network of transmission lines described by a graph $\mathcal{G} = (V, E)$. Then for each line $e \in E$ with spatial domain $[0, \ell_e]$ for some $\ell_e > 0$, we consider in $X_e \times X_e = L^2([0,\ell_e],\mathbb{R}^d) \times L^2([0,\ell_e],\mathbb{R}^d)$ the following system

$$\mathfrak{A}_e = \begin{bmatrix} -G_e & -D_e^* \\ -D_e^* & -R_e \end{bmatrix}, \quad \mathcal{H}_e(\xi) = \begin{bmatrix} C_e(\xi)^{-1} & 0 \\ 0 & L_e(\xi)^{-1} \end{bmatrix},$$
$$\mathcal{H}_e(\xi)x_e(t,\xi) = \begin{pmatrix} V_e(t,\xi) \\ I_e(t,\xi) \end{pmatrix}$$

and show that a boundary triplet is given

$$\Gamma_0^e(x_e) := \begin{bmatrix} V_e(0) \\ V_e(\ell_e) \end{bmatrix}, \Gamma_1^e(x) := \begin{bmatrix} \mathrm{i}I_e(0) \\ -\mathrm{i}I_e(\ell_e) \end{bmatrix}.$$

If we consider now the operators based on establishing continuity of Γ_0^e at every vertex $v \in V$, then this leads to the node-type boundary triplet

$$\Gamma_0^V((x_e)_{e\in E}) = (V(v))_{v\in V},$$
$$\Gamma_1^V((x_e)_{e\in E}) = \big\{\mathrm{i}\sum_{e\sim v,(e,k)\sim v} \operatorname{sgn}(e,v) I_e(k\ell_e)\big\}_{v\in V},$$

where the sum is taken over all edges e which are adjacent to v and $k = 0, 1$. Hence the voltage controlled boundary control system is given by

$$\dot{x} = \mathfrak{A}\mathcal{H}, \quad u(t) = \Gamma_0^V x, \quad y(t) = \Gamma_1^V x$$

where

$$x = (x_e)_{e\in E}, \quad \mathfrak{A} := \oplus_{e\in E}\mathfrak{A}_e,$$
$$\mathcal{H} := \oplus_{e\in E} \begin{bmatrix} C_e(\xi)^{-1} & 0 \\ 0 & L_e(\xi)^{-1} \end{bmatrix}$$

this can be viewed as having a voltage input at every node.

Our results will then be applied to show stability and the power balance equation for this interconnected system.

REFERENCES

Augner, B. (2020). Well-posedness and stability for interconnection structures of port-Hamiltonian type. In J. Kerner, H. Laasri, and D. Mugnolo (eds.), *Control Theory of Infinite-Dimensional Systems*, 1–52. Springer International Publishing, Cham.

Augner, B. and Jacob, B. (2014). Stability and stabilization of infinite-dimensional linear port-Hamiltonian systems. *Evolution Equations & Control Theory*, 3(2), 207.

Behrndt, J., Hassi, S., and De Snoo, H. (2020). *Boundary Value Problems, Weyl functions, and Differential Operators*. Springer Nature.

Egger, H., Kugler, T., Liljegren-Sailer, B., Marheineke, N., and Mehrmann, V. (2018). On structure-preserving model reduction for damped wave propagation in transport networks. *SIAM Journal on Scientific Computing*, 40(1), A331–A365. doi:10.1137/17M1125303.

Fiaz, S., Zonetti, D., Ortega, R., Scherpen, J., and van der Schaft, A. (2012). On port-Hamiltonian modeling of the synchronous generator and ultimate boundedness of its solutions. *IFAC Proceedings Volumes*, 45(19), 30–35. doi:https://doi.org/10.3182/20120829-3-IT-4022.00042. 4th IFAC Workshop on Lagrangian and Hamiltonian Methods for Non Linear Control.

Fiaz, S., Zonetti, D., Ortega, R., Scherpen, J.M.A., and van der Schaft, A.J. (2013). A port-Hamiltonian approach to power network modeling and analysis. *European Journal of Control*, 19(6), 477–485.

Jacob, B. and Zwart, H. (2012a). *Linear Port-Hamiltonian Systems on Infinite-dimensional Spaces*, volume 223 of *Operator Theory: Advances and Applications*. Birkhäuser.

Jacob, B. and Zwart, H. (2012b). *Linear port-Hamiltonian systems on infinite-dimensional spaces*. Operator Theory: Advances and Applications, 223. Birkhäuser/Springer Basel AG, Basel CH.

Kurula, M. and Zwart, H. (2015). Linear wave systems on n-d spatial domains. *International Journal of Control*, 88(5), 1063–1077. doi:10.1080/00207179.2014.993337.

Malinen, J. and Staffans, O. (2006). Conservative boundary control systems. *J. Differential Equations*, 231. doi:10.1016/j.jde.2006.05.012.

Malinen, J. and Staffans, O. (2007). Impedance passive and conservative boundary control systems. *Complex Anal. Oper. Theory*, 1. doi:10.1007/s11785-006-0009-3.

Philipp, F., Schaller, M., Faulwasser, T., Maschke, B., and Worthmann, K. (2021). Minimizing the energy supply of infinite-dimensional linear port-Hamiltonian systems. *IFAC-PapersOnLine*, 54(19), 155–160. doi:https://doi.org/10.1016/j.ifacol.2021.11.071. 7th IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control LHMNC 2021.

Ramirez, H., Le Gorrec, Y., Macchelli, A., and Zwart, H. (2014). Exponential stabilization of boundary controlled port-Hamiltonian systems with dynamic feedback. *IEEE Transactions on Automatic Control*, 59(10), 2849–2855.

Skrepek, N. (2021). *Linear port-Hamiltonian Systems on Multidimensional Spatial Domains*. Ph.D. thesis, Bergische Universität Wuppertal. doi:10.25926/g7h8-bd50.

Staffans, O. (2002). Passive and conservative continuous-time impedance and scattering systems. part i: Well-posed systems. *Math. Control Signals Systems*, 15, 291–315.

Trostorff, S. and Waurick, M. (2022). Characterisation for exponential stability of port-Hamiltonian systems. *arXiv:2201.10367*.

van der Schaft, A. and Stegink, T. (2016). Perspectives in modeling for control of power networks. *Annual Reviews in Control*, 41, 119–132. doi: https://doi.org/10.1016/j.arcontrol.2016.04.017.

Villegas, J. (2007). *A Port-Hamiltonian Approach to Distributed Parameter Systems*. Ph.D. thesis, University of Twente, Netherlands.

Villegas, J., Zwart, H., and Le Gorrec, Y.and Maschke, B. (2009). Exponential stability of a class of boundary control systems. *IEEE Transactions on Automatic Control*, 54. doi:10.1109/tac.2008.2007176.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Multi-modal behaviours in network SIR model

EXTENDED ABSTRACT

M. Alutto * **L. Cianfanelli** * **G. Como** * **F. Fagnani** *

* *Department of Mathematical Sciences, Politecnico di Torino*
(e-mail:{martina.alutto, leonardo.cianfanelli, giacomo.como, fabio.fagnani}@polito.it).

Abstract: We study the dynamical behaviour of the SIR network model at individual nodes. In two particular cases of a network consisting of only two nodes, we show how this behaviour differs from the epidemic outbreak and monotonic decreasing trend that occur in the scalar case. The first case deals with a network in which contact is only direct from one node to another, while the second treats a network with all contacts equal to one. The result shown for this scenario remains true by continuity for all those networks sufficiently close in parameter space.

Keywords: Nonlinear Systems and Control

1. INTRODUCTION

In recent years, there has been an increased focus on mathematical models that can be used to simulate, analyse and prevent the evolution of epidemic infections and identify parameters that can define patterns of behaviour. The classical deterministic SIR epidemic model, as first presented in the pioneering work Kermack and McKendrick (1927), is a compartmental model consisting of a nonlinear system of three coupled differential equations describing the evolution of the fractions of susceptible, infected, and recovered individuals in a fully mixed closed population. The main feature of this model is the existence of a phase transition described in terms of a scalar parameter, known as the reproduction number, whose value can determine two fundamentally different behaviours of the epidemics. Specifically, if the reproduction number does not exceed 1, then the fraction of the infected individuals remains monotonically decreasing in time, and thus preventing an epidemic outbreak. In contrast, if the reproduction number exceeds the unitary threshold, then the fraction of infected individuals is initially increasing until reaching a peak, after which it starts to decrease monotonically and vanishes asymptotically in the large time limit. Most compartmental models of disease propagation, such as in Hethcote (2000), assume that populations are fully mixed, meaning that an infected individual is equally likely to spread the infection to any other member of the population. However, this assumption is unrealistic, as one infected individual is not equally likely to infect all others because in the real world each individual has contact with only a small fraction of the total population. In order to better describe a disease spread, several studies have considered topological properties of various networks and studied their effects on the epidemic processes taking place on these Newman (2002). Through the study of network SIR models, the topology of the graph has proved to be a determinant feature for the dynamics of the system, as in Mei et al. (2017) where a new threshold condition for epidemic behaviour is proposed in terms of network characteristics, initial conditions and infection parameters. It has been shown how a weighted average of the population of infected people is able to capture all the information on epidemic evolution, where the weights depend on network properties.

However, in individual nodes the dynamics may present different phenomena and our contribution concerns precisely the investigation of such scenarios. We will focus on two specific networks, in the first case there is a direct contact only from one node to another, while in the second case we will consider a particular network where the contact matrix has all entries equal to one.

2. NETWORK SIR MODEL

In this section, we introduce the network SIR model. Let us consider a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ with finite set of nodes $\mathcal{V} = \{1, 2, \dots, n\}$, set of directed links $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and a matrix A in $\mathbb{R}_+^{n\times n}$, whose entries embody the strength of both the infection and the contact frequency of members of subpopulation i with members of subpopulation j. For given recovery rate $\gamma > 0$, the network SIR epidemic model on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ is the dynamical system

$$\begin{cases} \dot{x}_i = -x_i \sum_{j=1}^{n} a_{ij} y_j \\ \dot{y}_i = x_i \sum_{j=1}^{n} a_{ij} y_j - \gamma y_i \\ \dot{z}_i = \gamma y_i \end{cases} \tag{1}$$

for $i = 1, \dots, n$, where x_i, y_i, and z_i represent respectively the fractions of susceptible, infected, and recovered individuals in population i. Notice that (1) can be more compactly rewritten in its vectorial form

$$\begin{cases} \dot{x} = -diag(x)Ay \\ \dot{y} = diag(x)Ay - \gamma y \end{cases} \tag{2}$$

This model has been studied in Mei et al. (2017) and Nowzari et al. (2016). In particular, they show that if $y(0) > \mathbf{0}_n$, and $x(0) \geq \mathbf{0}_n$, then $t \mapsto x(t)$ and $t \mapsto y(t)$ are strictly positive for all $t \geq 0$. The strictly positivity of the solutions will be useful in the following theorem. It has also been proven that all solutions converge to an equilibrium point of the form $(x^*, 0, z^*) \in \mathbb{R}^{3n}_+$ such that $x^* + z^* = \mathbf{1}$ and that the locally asymptotically stable equilibrium points are those such that

$$\lambda_{\max}(\mathrm{diag}\,(x^*)A) < \gamma$$

where $\lambda_{\max}$ is the dominant eigenvalue of the nonnegative matrix $\mathrm{diag}\,(x^*)A$, which coincides with its spectral radius thanks to the Perron-Frobenius Theorem. Under the assumption of strong connectivity of the graph $\mathcal{G}$, (Mei et al., 2017, Theorem 7) shows that the quantity

$$R(t) = \lambda_{\max}(\mathrm{diag}\,(x(t))A)/\gamma$$

is decreasing along solutions and it plays a role similar to the one played by the reproduction number in the scalar SIR model. Specifically, if $R(0) \leq 1$ then the weighted average of the infected $v(0)'y(t)$ will be monotonically decreasing to 0 as t grows large, where $v(t)$ stands for the corresponding left-eigenvector of λ_{max} of the matrix $\mathrm{diag}\,(x(t))A$. On the other hand, if $R(0) > 1$, then the weighted average $v(0)'y(t)$ will be initially increasing (epidemic outbreak) and there exists some $\tau > 0$ such that $R(\tau) \leq 1$ and the weighted average $v(\tau)'y(t)$ will be decreasing to 0 for t in the interval $[\tau, +\infty)$. It is therefore possible to find a parallelism with the scalar case.

However, several simulations show how the dynamical behaviour at individual nodes can exhibit atypical phenomena compared to the unimodal or monotonically decreasing behaviour of the scalar case. We will investigate these phenomena, trying to determine which conditions are sufficient for them to occur and which are necessary to have a behaviour similar to the classical SIR case at each node. For this analysis from now on, we will focus on the case of a network consisting of two nodes.

3. TWO-NODES NETWORK

In this section, we will deal with the case of a network composed of two nodes in two different scenarios: in the first case, the contact only occurs directly from the second node to the other and not vice versa, while in the second one, we will consider an homogeneous network in which the contact matrix has all entries equal to 1.

3.1 One node cannot be infected by the other one

Consider the SIR model (1) in the case of a network composed of only two nodes, with the following infection/weight matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix} \tag{3}$$

In this scenario, the second node is isolated from the first in the sense that the individuals of the second node cannot be infected due to contacts with individuals of the first node. On the other hand, in the first node, the infection process occurs both due to infection within the node, and to infection received through contact with the second node. As the second node is unaffected by contact with the first, its dynamics is equivalent to scalar SIR dynamics. Therefore, if $a_{22}x_2(0) > \gamma$, then $t \mapsto y_2(t)$ is initially monotonically increasing, it shows an epidemic outbreak and then it becomes exponentially decreasing to 0. We assume for simplicity of notation $z(0) = \mathbf{0}$, but the results can be extended to the general case without loss of generality.

Let t_2 denote the time at which the peak of the second node occurs. We can derive t_2 from Harko et al. (2014), by using the fact that $z_2(0) = \mathbf{0}$ and $x_2(t_2) = \gamma/a_{22}$. Specifically,

$$t_2 = \int_1^{\frac{\gamma}{a_{22}x_2(0)}} \frac{ds}{s(-a_{22} - \gamma log(s) + a_{22}x_2(0)s)} \tag{4}$$

while the infected at the second node peak will be

$$y_2(t_2) = \frac{\gamma}{a_{22}} \ln\left(\frac{\gamma}{a_{22}x_2(0)}\right) - \frac{\gamma}{a_{22}x_2(0)} + 1. \tag{5}$$

Let us analyse the dynamics of the first node in this scenario.

Theorem 1. Consider the SIR model (1) in the case of a network composed of only two nodes with matrix (3). Suppose $(x_i(0), y_i(0), z_i(0))$ for $i = 1, 2$ as initial condition such that the following relations hold:

$$a_{11}x_1(0)y_1(0) + a_{12}x_1(0)y_2(0) - \gamma y_1(0) < 0 \tag{6}$$

$$(a_{22}x_2(0) - \gamma)y_2(0) > 0 \tag{7}$$

$$(x_1(0) - \gamma y_1(0)t_2)a_{12}y_2(t_2) > \gamma y_1(0) \tag{8}$$

where t_2 is the time instant at which the infected curve of the second node has the infection peak. Then,

a) $\dot{y}_1(0) < 0$ and $\dot{y}_2(0) > 0$

b) $\exists t^* \leq t_2$ such that $\dot{y}_1(t^*) > 0$.

Proof. The first statement is obvious from the relations in (6)-(7). Let us now assume by contradiction that the statement *b*) is false. This would mean that for $t \in [0, t_2]$,

$$\dot{y}_1(t) \leq 0 \quad \Rightarrow \quad y_1(t) \leq y_1(0) \tag{9}$$

From fundamental theorem of integral calculus, we can observe that

$$x_1(t_2) = x_1(0) + \int_0^{t_2} \dot{x}_1(s)ds \tag{10}$$

Since for all $t \in [0, t_2]$,

$$\dot{y}_1(t) = x_1(t)(a_{11}y_1(t) + a_{12}y_2(t)) - \gamma y_1(t) \tag{11}$$

$$= -\dot{x}_1(t) - \gamma y_1(t) \leq 0, \tag{12}$$

then $\dot{x}_1(t) \geq -\gamma y_1(t) \geq -\gamma y_1(0)$. Therefore

$$x_1(t_2) \geq x_1(0) - \int_0^{t_2} \gamma y_1(s)ds \geq x_1(0) - \gamma y_1(0)t_2. \tag{13}$$

We can state that

$$\dot{y}_1(t_2) = x_1(t_2)(a_{11}y_1(t_2) + a_{12}y_2(t_2)) - \gamma y_1(t_2) \tag{14}$$

$$> x_1(t_2)a_{12}y_2(t_2)) - \gamma y_1(t_2) \tag{15}$$

$$\geq (x_1(0) - \gamma y_1(0)t_2)a_{12}y_2(t_2) - \gamma y_1(t_2) \tag{16}$$

$$\geq \gamma y_1(0) - \gamma y_1(t_2) \geq 0 \tag{17}$$

where the first inequality follows from the strict positivity of $y(t)$, while the last one is true for (9). This is an absurd and therefore there must be a time instant $t^* \leq t_2$ such that the first derivative of y_1 becomes positive.

□

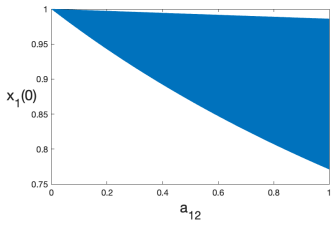


Fig. 1. Values of initial condition of the first node $x_1(0)$ as a_{12} varies, with the other parameters fixed.

Remark 1. Note that the theorem hypothesis of initial decreasing trend of y_1 and the validity of the relation (8) are not incompatible. In fact, if we decide to set the parameters $\gamma = 0.9$, $a_{22} = 1$, $a_{11} = 0.1$ and the initial condition of the second node $x_2(0) = 0.99$, we find that when a_{12} varies, the values of $x_1(0)$ that are admissible by the theorem are those shown in the Figure (1).

Remark 2. The theorem shows how an atypical behaviour can occur in this particular scenario. The curve of infected in the first node, instead of decreasing as it would if it were isolated, causes an outbreak of the epidemic with a second peak of infection (the first one is in initial condition) and then returns to decrease.

Let us now study a sufficient condition for this phenomenon not to occur.

Theorem 2. Consider the SIR model (1) in the case of a network composed of only two nodes with matrix (3). Suppose $(x_i(0), y_i(0), z_i(0))$ for $i = 1, 2$ as initial condition such that $x_2(0) < (a_{12}+\gamma)/(a_{12}+a_{22})$. Then, the following statements hold:

a) Define the set
$$\Omega = \{(x_1, y_1, x_2, y_2) \in [0,1]^4 : \dot{y}_1 < 0\} \tag{18}$$
then Ω is an invariant set for the dynamics,

b) For any initial condition such that $(x_1(0), y_1(0), y_2(0)) \notin \Omega$, there exists an instant $t^* > 0$:
$$(x_1(t), y_1(t), x_2(t), y_2(t)) \notin \Omega \quad \forall t < t^* \tag{19}$$
$$(x_1(t), y_1(t), x_2(t), y_2(t)) \in \Omega \quad \forall t > t^* \tag{20}$$

Proof. Regarding statement 3), consider initial conditions such that $(x_1(0), y_1(0), x_2(0), y_2(0)) \in \Omega$ and let $(x_i(t), y_i(t))$ be the corresponding solutions in node i.
We want to prove that $(x_1(t), y_1(t), x_2(t), y_2(t)) \in \Omega \quad \forall t$, i.e. if the first node has initial condition that make its curve of infected decreasing, then this curve will remain decreasing $\forall t \geq 0$. Suppose that exist an instant $t^* > 0$ such that
$$(x_1(t^*), y_1(t^*), x_2(t^*), y_2(t^*)) \in \delta\Omega, \tag{21}$$
where $\delta\Omega = \{(x_1, y_1, x_2, y_2) \in [0,1]^4 : \dot{y}_1 = 0\}$ and $(x_1(t), y_1(t), x_2(t), y_2(t)) \in \Omega \quad \forall t < t^*$.
Let us consider the first derivative of $\dot{y}_1$ with respect to time t.
$$\ddot{y}_1 = \dot{x}_1(a_{11}y_1 + a_{12}y_2) + x_1(a_{11}\dot{y}_1 + a_{12}\dot{y}_2) - \gamma\dot{y}_1 \tag{22}$$
Notice that, necessarily, $y_1(t^*) > 0$ and $\dot{y}_1(t^*) = 0$. We can observe that at t^*
$$\ddot{y}_1(t^*) = \dot{x}_1(a_{11}y_1 + a_{12}y_2) + a_{12}x_1\dot{y}_2 \tag{23}$$
$$= -x_1(a_{11}y_1 + a_{12}y_2)^2 + a_{12}x_1(a_{22}x_2 - \gamma)y_2 \tag{24}$$
If $a_{22}x_2(t^*) < \gamma$, then this quantity is negative.

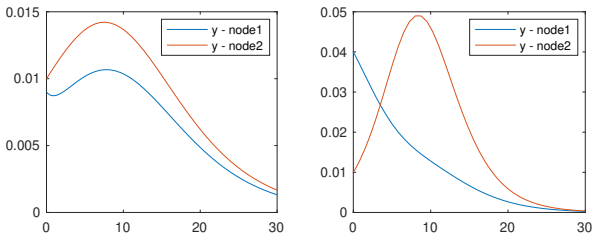


Fig. 2. Simulations of the network SIR model. The scenario on the left is under the assumptions of Theorem 1 and y_1 has a second peak of infection, while the one on the right meets the Theorem 2' assumptions and no multimodal behaviour occurs.

Otherwise, y_2 is still increasing and we have
$$\begin{aligned}\ddot{y}_1(t^*) &= -x_1(a_{11}y_1 + a_{12}y_2)^2 + a_{12}x_1(a_{22}x_2 - \gamma)y_2 \\ &< x_1y_2(-a_{12}^2y_2 + a_{12}a_{22}x_2 - a_{12}\gamma) \\ &< a_{12}x_1y_2(-a_{12}y_2(0) + a_{22}x_2(0) - \gamma) < 0\end{aligned}$$
where the first inequality holds because $a_{12}y_2(t) \leq a_{11}y_1(t) + a_{12}y_2(t)$ for all t, while the second one follows from the monotonically decreasing trend of x_2 and the still increasing behaviour of y_2. The final inequality is a direct consequence of the relation on the initial condition.
Hence, $\ddot{y}_1(t^*) < 0$ and, by continuity, we can state that $\ddot{y}_1 < 0$ in a neighborhood $(t^* - \epsilon, t^*]$. Since $\dot{y}_1$ is decreasing and, by construction, $\dot{y}_1 < 0$ for $t < t^*$, it is not possible that $\dot{y}_1(t^*) = 0$ and this proves the statement *a*).

For statement *b*), having shown the previous point, we only have to prove that exists an instant t^* such that $(x_1(t^*), y_1(t^*), x_2(t^*), y_2(t^*)) \in \Omega$. If this was not the case, then it would be $\dot{y}_1(t) > 0 \, \forall t$ and $y_1(t)$ would not tend to 0. Therefore $z_1(t)$ would blow up which is impossible for the first statement. □

In Figure 2, we compare the different behaviours of the infected curves in the two cases considered in the previous theorems.

3.2 Homogeneous network with $A = \beta\mathbf{1}\mathbf{1}'$

Let us now analyse a network in which the contact-infection matrix $A = \beta\mathbf{1}\mathbf{1}'$, where β is the explicit infection rate and with no distinction between individuals belonging to the same or different populations. We will focus on a particular scenario in which one node has an initial small fraction of infected, while the other one is initially totally susceptible. Also in this case, we find some initial conditions that ensure the occurrence of an atypical behaviour. The result shown is valid by continuity for all those sufficiently close in parameter space.

Theorem 3. Consider the network SIR model (1), complete contact graph with $A = \mathbf{1}\mathbf{1}'$, and unitary infection and recovery rates $\beta = \gamma = 1$. Let the initial condition
$$x_1(0) = 1 - \varepsilon\,, \quad y_1(0) = \varepsilon\,, \quad x_2(0) = 1\,, \quad y_2(0) = 0\,,$$
for some $0 < \varepsilon < \bar{\varepsilon}$, where
$$\bar{\varepsilon} = \min\{\varepsilon \in [0,1] : \varepsilon = \frac{1-\varepsilon}{2-\varepsilon}\Big(1 - \ln(2-\varepsilon)\Big)\}\,.$$
Then, the fraction of infected in the first population y_1 changes monotonicity exactly twice along the solution.

Proof. The dynamics is given by

$$\dot{x}_i = -2x_i\overline{y}\,, \qquad \dot{y}_i = 2x_i\overline{y} - y_i\,, \qquad i = 1, 2\,,$$

where

$$\overline{x} = (x_1 + x_2)/2\,, \qquad \overline{y} = (y_1 + y_2)/2\,,$$

are the mean aggregate variables. Notice that

$$\dot{\overline{x}} = -2\overline{x}\,\overline{y}\,, \qquad \dot{\overline{y}} = (2\overline{x} - 1)\overline{y}\,,$$

so that by scalar theory the quantity $\overline{x}+\overline{y}-\frac{1}{2}\ln\overline{x}$ remains constant along solutions, i.e. $\forall t \geq 0$,

$$\overline{x}(t) + \overline{y}(t) - \frac{1}{2}\ln\overline{x}(t) = \overline{x}(0) + \overline{y}(0) - \frac{1}{2}\ln\overline{x}(0)\,. \quad (25)$$

On the other hand, whenever $x_2 > 0$, we have

$$\left(\frac{\dot{x}_1}{x_2}\right) = \frac{\dot{x}_1 x_2 - x_1\dot{x}_2}{x_2^2} = 0$$

so the ratio x_1/x_2 remains constant along solutions, i.e.,

$$x_i(t)\overline{x}(0) = x_i(0)\overline{x}(t)\,, \qquad \forall t \geq 0\,,\ i = 1, 2\,. \quad (26)$$

Observe that

$$\dot{y}_1(0) = 2x_1(0)\overline{y}(0) - y_1(0) = -\varepsilon^2 < 0\,,$$

whereas

$$\dot{\overline{y}}(0) = (2\overline{x}(0) - 1)\overline{y}(0) = (1 - \varepsilon)\varepsilon/2 > 0\,.$$

Let $t^* > 0$ be the aggregate peak time, i.e., the time such that

$$\overline{x}(t^*) = 1/2, \qquad \dot{\overline{y}}(t^*) = 0\,.$$

It follows from (25) that

$$\overline{y}(t^*) = \overline{x}(0) + \overline{y}(0) - \frac{1}{2}\ln\overline{x}(0) - \overline{x}(t^*) + \frac{1}{2}\ln\overline{x}(t^*) \quad (27)$$

$$= \frac{1}{2}\Big(1 - \ln(2 - \varepsilon)\Big)\,, \quad (28)$$

while (26) implies that

$$x_1(t^*) = \frac{x_1(0)\overline{x}(t^*)}{\overline{x}(0)} = \frac{1 - \varepsilon}{2 - \varepsilon}\,. \quad (29)$$

Now, assume by contradiction that

$$\dot{y}_1(t) \leq 0\,, \qquad \forall t \geq 0\,. \quad (30)$$

In particular, this would in particular imply that

$$\begin{aligned} 0 \geq \dot{y}_1(t^*) &= 2x_1(t^*)\overline{y}(t^*) - y_1(t^*) \\ &= \frac{1 - \varepsilon}{2 - \varepsilon}\Big(1 - \ln(2 - \varepsilon)\Big) - y_1(t^*) \\ &\geq \frac{1 - \varepsilon}{2 - \varepsilon}\Big(1 - \ln(2 - \varepsilon)\Big) - y_1(0) \end{aligned}$$

So we would obtain that

$$y_1(0) = \varepsilon \geq \frac{1 - \varepsilon}{2 - \varepsilon}\Big(1 - \ln(2 - \varepsilon)\Big)$$

Therefore necessarily $\varepsilon \geq \bar{\varepsilon}$, thus violating the assumption. It then follows that there exists at least a time $\bar{t}$ such that $\dot{y}_1(\bar{t}) > 0$. The claim then follows, since $\dot{y}_1(0) < 0$ and $\lim_{t\to+\infty} y_1(t) = 0$. We have shown that the infected curve in the first node y_1 changes monotonicity at least twice along the solution. □

Based on the previous theorem and standard continuity arguments we can prove the following result.

Proposition 1. There exist values $\bar{\varepsilon} > 0$, $\underline{\beta} < 1 < \overline{\beta}$, and $\underline{\gamma} < 1 < \overline{\gamma}$, and an open subset of nonnegative matrices $\mathcal{M} \subseteq \mathbb{R}^{2\times 2}$ containing $\mathbf{1}\mathbf{1}'$ such that the network SIR model with $n = 2$ subpopulations, contact graph with weight matrix $A \in \mathcal{M}$, infection rate $\beta \in (\underline{\beta}, \overline{\beta})$, recovery rate $\gamma \in (\underline{\gamma}, \overline{\gamma})$ with any initial condition such that

$$0 < y_i(0) = 1 - x_i(0) < \bar{\varepsilon}\,, \qquad i = 1, 2\,,$$

has solution such that the faction of infected in the first population $y_1(t)$ is a multi-modal function of t.

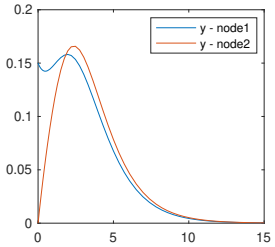


Fig. 3. Simulation of the SIR model on an homogeneous network where the second node is initially totally susceptible.

In Figure 3, we can observe the occurrence of this atypical behaviour also in this scenario: the curve of the infected in the first node is initially decreasing, then it becomes increasing to a peak of infection, and then decreasing to 0. We have considered a population characterized by a small initial fraction of infected people, such as to cause an exponential decrease to 0, using the scalar SIR model theory. However, if this first population meets a totally healthy one, it begins to infect it and then suffers a return wave of infection with a second peak (the first one was in its initial condition).

4. CONCLUSION

In this work, we analyse the behaviour of the network SIR model at each node in some particular cases and define sufficient conditions for the occurrence of an atypical phenomenon, compared to the classical SIR theory results. This phenomenon consists of the appearance of multi-modal dynamical behaviours at a single node and it has been shown in different scenarios of a two-nodes network.

REFERENCES

Harko, T., Lobo, F.S.N., and Mak, M.K. (2014). Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Appl. Math. Comput.*, 236, 184–194.

Hethcote, H.W. (2000). The mathematics of infectious diseases. *SIAM Review*, 42(4), 599–653.

Kermack, W.O. and McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772), 700–721.

Mei, W., Mohagheghi, S., Zampieri, S., and Bullo, F. (2017). On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, 44, 116–128.

Newman, M.E.J. (2002). Spread of epidemic disease on networks. 66(1).

Nowzari, C., Preciado, V.M., and Pappas, G.J. (2016). Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems Magazine*, 36(1), 26–46.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Towards Funnel MPC for nonlinear systems with relative degree two ⋆

Dario Dennstädt *

* *Technische Universität Ilmenau, Weimarer Str. 25, 98693 Ilmenau, Germany (dario.dennstaedt@tu-ilmenau.de).*

Abstract: Funnel MPC, a novel Model Predictive Control (MPC) scheme, allows guaranteed output tracking of smooth reference signals with prescribed error bounds for nonlinear multi-input multi-output systems. To this end, the stage cost resembles the high-gain idea of funnel control. Without imposing additional output constraints or terminal conditions, the Funnel MPC scheme is initially and recursively feasible for systems with relative degree one and stable internal dynamics. Using an additional funnel for the derivative as a penalty term in the stage cost, these results can be also extended to single-input single-output systems with relative degree two.

Keywords: model predictive control, funnel control, output tracking, nonlinear systems

1. INTRODUCTION

Model Predictive Control (MPC) is a widely-used control technique for linear and nonlinear systems and has seen various applications, see e.g. Qin and Badgwell (2003). Key reasons for its success are its applicability to multi-input multi-output nonlinear systems and its ability to directly take control and state constraints into account. To this end, a finite-horizon Optimal Control Problem (OCP) is solved before the prediction horizon is shifted forward in time and the procedure is repeated ad infinitum, see e.g. Grüne and Pannek (2017) and Coron et al. (2020).

Recursive feasibility is essential for successfully applying MPC. This means, solvability of the OCP at a particular time instant has to automatically imply solvability of the OCP at the successor time instant. In order to achieve this, often, suitably designed terminal conditions (cost and constraints) are incorporated in the OCP to be solved at each time instant, see Rawlings et al. (2017). However, such (artificially introduced) terminal conditions increase the computational burden of solving the OCP and complicate the task of finding an initially-feasible solution. As a consequence, the domain of the MPC feedback controller might become significantly smaller, see e.g. Chen et al. (2003); González and Odloak (2009). This technique becomes considerably more involved in the presence of time-varying state constraints, see e.g. Manrique et al. (2014).

To overcome these restrictions, Funnel MPC (FMPC) was proposed in Berger et al. (2020). This allows output tracking such that the tracking error evolves in a pre-specified, potentially time-varying performance funnel. A "funnel-like" stage cost, which penalizes the tracking error and becomes infinite when approaching the funnel boundary, is used. By incorporating output constraints in the OCP and using properties of the system class in consideration, initial and recursive feasibility are shown – without imposing additional terminal conditions and independent of the length of the prediction horizon. The novel stage cost used in FMPC is inspired by funnel control, a model-free output-error feedback controller first proposed in Ilchmann et al. (2002). The funnel controller is an adaptive controller which allows output tracking within a prescribed performance funnel for a fairly large class of systems solely invoking structural assumptions, i.e. stable internal dynamics, known relative degree, and a sign-definite high-frequency gain matrix.

It is shown in Berger et al. (2021) that such funnel-inspired stage cost automatically ensure initial and recursive feasibility for a class of nonlinear systems with relative degree one and, in a certain sense, input-to-state stable internal dynamics. Since the requirement of a sign-definite gain matrix is omitted, the system class is larger than the one the original funnel controller is applicable to. Moreover, adding (artificial) output constraints to the OCP, as used in Berger et al. (2020), is superfluous. In numerical simulations, FMPC shows superior performance compared to both MPC with quadratic stage cost and funnel control.

Utilizing so-called feasibility constraints in the OCP and restricting the class of admissible funnel functions, the findings in Berger et al. (2021) are generalized to systems with arbitrary relative degree in Berger and Dennstädt (2022). We show that for single-input single-output systems with relative degree two the addition of such constraints to the Optimal Control Problem is not necessary. However, while previous results allow for an arbitrary short prediction horizon, a sufficiently long horizon – depending on the funnel – is necessary.

2. SYSTEM CLASS AND CONTROL OBJECTIVE

We consider control affine single-input single-output systems

⋆ D. Dennstädt gratefully thanks the Technische Universität Ilmenau and the Free State of Thuringia for their financial support as part of the Thüringer Graduiertenförderung.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + g(x(t))u(t), \quad x(t^0) = x^0, \\ y(t) &= h(x(t)), \end{aligned} \tag{1}$$

with $t^0 \in \mathbb{R}_{\geq 0}$, $x^0 \in \mathbb{R}^n$, functions $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R}^n)$, $g \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R}^{n})$, $h \in \mathcal{C}^3(\mathbb{R}^n, \mathbb{R})$ and a control function $u \in L^\infty_{\mathrm{loc}}(\mathbb{R}_{\geq 0}, \mathbb{R})$. The system (1) has a *solution* in the sense of *Carathéodory*, that is an absolutely continuous function $x : [t^0, \omega) \to \mathbb{R}^n$, $\omega > t^0$, with $x(t^0) = x^0$ which satisfies the ODE in (1) for almost all $t \in [t^0, \omega)$.

We recall the notion of relative degree for system (1). Assuming that f, g, h are sufficiently smooth, the Lie derivative of h along f is defined by $(L_f h)(x) := h'(x) f(x)$. Lie derivatives of higher order are recursively defined by $L_f^k h := L_f(L_f^{k-1} h)$, for $k \in \mathbb{N}$, with $L_f^0 h = h$. Then system (1) is said to have *(global and strict) relative degree* $r \in \mathbb{N}$, if $\forall\, k \in \{1, \dots, r-1\}\ \forall\, x \in \mathbb{R}^n$:

$$(L_g L_f^{k-1} h)(x) = 0 \quad \text{and} \quad (L_g L_f^{r-1} h)(x) \neq 0.$$

If (1) has relative degree r, then, under the additional assumptions provided in (Byrnes and Isidori, 1991, Cor. 5.6), there exists a diffeomorphic coordinate transformation

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n, \Phi(x(t)) = (y(t), \dot{y}(t), \dots, y^{(r-1)}(t), \eta(t)) \tag{2}$$

which puts the system into Byrnes-Isidori form

$$\begin{aligned} y^{(r-1)}(t) &= p\big(y(t), \dot{y}(t), \dots, y^{(r-1)}(t), \eta(t)\big) \\ &\quad + \gamma\big(y(t), \dot{y}(t), \dots, y^{(r-1)}(t), \eta(t)\big)\, u(t), \quad \text{(3a)} \\ \dot{\eta}(t) &= q\big(y(t), \dot{y}(t), \dots, y^{(r-1)}(t), \eta(t)\big), \quad \text{(3b)} \end{aligned}$$

where $p \in \mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$, $q \in \mathcal{C}^0(\mathbb{R}^n, \mathbb{R}^{n-r})$, $\gamma \in \mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$ and $(y(t^0), \dot{y}(t^0), \dots, y^{(r-1)}(t^0), \eta(t^0)) = \Phi(x^0)$. Furthermore, we require the following *bounded-input, bounded-state* (BIBS) condition on the internal dynamics (3b):

$$\begin{aligned} &\forall\, c_0 > 0\ \exists\, c_1 > 0\ \forall\, t^0 \geq 0\ \forall\, \eta^0 \in \mathbb{R}^{n-r} \\ &\quad \forall\, \zeta \in L^\infty_{\mathrm{loc}}([t^0, \infty), \mathbb{R}^{rm}) : \ \big\|\eta^0\big\| + \|\zeta\|_\infty \leq c_0 \\ &\qquad \Longrightarrow \ \big\|\eta(\cdot; t^0, \eta^0, \zeta)\big\|_\infty \leq c_1, \end{aligned} \tag{4}$$

where $\eta(\cdot; t^0, \eta^0, \zeta) : [t^0, \infty) \to \mathbb{R}^{n-r}$ denotes the unique global solution of (3b) when $(y, \dots, y^{(r-1)})$ is substituted by ζ. The maximal solution $\eta(\cdot; t^0, \eta^0, \zeta)$ can indeed be extended to a global solution due to the BIBS condition (4).

Throughout this note we will assume that the system (1) has relative degree $r = 2$ and that there exists a diffeomorphism Φ as in (2) which puts the system into the Byrnes-Isidori form (3).

The objective is to design a control strategy which allows the output tracking of given reference trajectories $y_{\mathrm{ref}} \in W^{2,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R})$ within pre-specified error bounds. To be precise, the tracking error $t \mapsto e(t) := y(t) - y_{\mathrm{ref}}(t)$ and its derivative $\dot{e}(t)$ shall evolve within the prescribed performance funnels

$$\mathcal{F}_{\psi_i} := \{\, (t, e) \in \mathbb{R}_{\geq 0} \times \mathbb{R} \mid \|e\| < \psi_i(t) \,\}, \quad i = 0, 1,$$

see also Figure 1. These funnels are determined by the choice of the functions ψ_0, ψ_1 belonging to

$$\mathcal{G}^0 := \left\{ \psi \in W^{2,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R}) \;\middle|\; \inf_{t \geq 0} \psi(t) > 0 \right\}.$$

Note that the funnel ψ_i is uniformly bounded away from zero; i.e. there exists a boundary $\lambda > 0$ with $\psi_i(t) > \lambda$ for all $t \geq 0$. Thus, perfect or asymptotic tracking is not our control objective. However, λ can be chosen arbitrarily small.

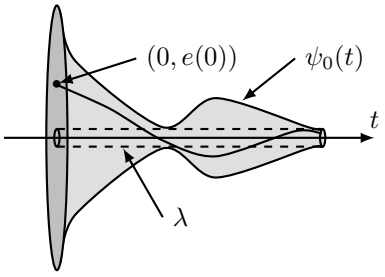


Fig. 1. Error evolution in a funnel $\mathcal{F}_{\psi_0}$ with boundary ψ_0.

If the error e evolves within the funnel $\mathcal{F}_{\psi_0}$ for some $\psi_0 \in \mathcal{G}^0$, then the derivative $\dot{e}$ has to satisfy at some point $t \geq 0$

$$\dot{e}(t) \leq \dot{\psi}_0(t) \text{ or } \dot{e}(t) \geq -\dot{\psi}_0(t).$$

Thus, the derivative funnel must be large enough for the error e to follow the funnel boundary ψ_0 and we therefore assume that $\psi = (\psi_0, \psi_1)$ is an element of

$$\mathcal{G}^1 := \left\{ (\psi_0, \psi_1) \in \mathcal{G}^0 \times \mathcal{G}^0 \;\middle|\; \exists\, \varepsilon > 0\, \forall\, t \geq 0 : \psi_1(t) \geq \varepsilon - \dot{\psi}_0(t) \right\}.$$

Typically, the specific application dictates constraints on the tracking error and thus indicates suitable choices for ψ.

3. FUNNEL MPC

In order to extend the results from Berger et al. (2021) to systems of the form (1) with relative degree two, we define, for $y_{\mathrm{ref}} \in W^{2,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R})$, $t \geq 0$, and $\zeta = (\zeta^0, \zeta^1) \in \mathbb{R}^2$,

$$e_i(t, \zeta) := \zeta^i - y_{\mathrm{ref}}^{(i)}(t) \quad \text{for } i = 0, 1.$$

We propose, for $\psi = (\psi_0, \psi_1) \in \mathcal{G}^1$ and the design parameter $\lambda_u \geq 0$, the new *stage cost function*

$\ell : \mathbb{R}_{\geq 0} \times \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$,

$$(t, \zeta, u) \mapsto \begin{cases} \sum_{i=0}^{1} \frac{1}{1 - \|e_i(t,\zeta)\|^2 / \psi_i(t)^2} + \lambda_u \|u\|^2, & \|e_i(t,\zeta)\| \neq \psi_i(t) \\ & \text{for } i = 0, 1 \\ \infty, & \text{else.} \end{cases} \tag{5}$$

By setting $\zeta = (y(t), \dot{y}(t))$, the terms $\frac{1}{1 - \|e_i(t,\zeta)\|^2 / \psi_i(t)^2}$ penalize the distance of the tracking error $e(t) = y(t) - y_{\mathrm{ref}}(t)$ and its derivative $\dot{e}(t)$ to their respective funnel boundaries $\psi_i(t)$. The parameter λ_u allows to adjust a suitable trade off between tracking performance and required control effort. The stage cost ℓ is motivated by the design of the funnel controller in Hackl et al. (2013) which also introduces an additional funnel for the derivative in order to generalize the results from Ilchmann et al. (2002) to systems with relative degree two.

Based on the stage cost (5), we define the Funnel MPC (FMPC) algorithm as follows.

Algorithm 1. (FMPC).
Given: System (1), reference signal $y_{\mathrm{ref}} \in W^{2,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R})$, funnel function $\psi = (\psi_0, \psi_1) \in \mathcal{G}^1$, stage cost function ℓ as in (5), $M > 0$, $t^0 \in \mathbb{R}_{\geq 0}$, and $x^0 \in \mathbb{R}^n$ with

$$\Phi(x^0) \in \mathcal{D}_{t^0} := \left\{ (\zeta, \eta) \in \mathbb{R}^2 \times \mathbb{R}^{n-2} \;\middle|\; \begin{matrix} \|e_i(t_0, \zeta)\| < \psi_i(t_0) \\ \text{for } i = 0, 1 \end{matrix} \right\},$$

where Φ is the diffeomorphism from (2).
Set the time shift $\delta > 0$, the prediction horizon $T > \delta$ and initialize the current time $\hat{t} := t^0$.
Steps:

(a) Obtain a measurement of the state $x = \Phi^{-1}(y, \dot{y}, \eta)$ at time $\hat{t}$ and set $\hat{x} := x(\hat{t})$.

(b) Compute a solution $u^\star \in L^\infty([\hat{t}, \hat{t}+T], \mathbb{R})$ of the Optimal Control Problem (OCP)

$$\begin{aligned} \underset{u \in L^\infty([\hat{t},\hat{t}+T],\mathbb{R})}{\text{minimize}} \quad & \int_{\hat{t}}^{\hat{t}+T} \ell\big(t, (y(t), \dot{y}(t)), u(t)\big) \mathrm{d}t \\ \text{subject to} \quad & (1),\ x(\hat{t}) = \hat{x}, \\ & \|u(t)\| \leq M \quad \text{for } t \in [\hat{t}, \hat{t}+T] \end{aligned} \tag{6}$$

(c) Apply the feedback law

$$\mu : [\hat{t}, \hat{t}+\delta) \times \mathbb{R}^n \to \mathbb{R}, \quad \mu(t, \hat{x}) = u^\star(t) \tag{7}$$

to system (1). Increase $\hat{t}$ by δ and go to Step (a).

In practical application there usually is a limitation $M > 0$ on the maximal control that can be applied to the system 1. The constraint $\|u(t)\| \leq M$ in the OCP (6) ensures that the control signal meets this bound.

The following theorem shows that for systems with relative degree two the Funnel MPC Algorithm 1 is, given a sufficiently long prediction horizon $T > 0$ and large enough control constraint $M > 0$, initially and recursively feasible and that it guarantees the evolution of the tracking error e and its derivative $\dot{e}$ within their respective performance funnels $\mathcal{F}_{\psi_i}$. Due to space limitations, we omit the proof.

Theorem 2. Consider system (3) with strict relative degree $r = 2$ and assume that there exists a diffeomorphism $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ such that the coordination transformation in (2) puts the system (1) in the Byrnes-Isidori form (3) satisfying (4). Let $\psi = (\psi_0, \psi_1) \in \mathcal{G}^1$, $y_{\mathrm{ref}} \in W^{2,\infty}(\mathbb{R}_{\geq 0}, \mathbb{R})$, $t^0 \in \mathbb{R}_{\geq 0}$, $\delta > 0$, and $B \subset \mathcal{D}_{t^0}$ a compact set. Then there exist $T > \delta$ and $M > 0$ such that the FMPC Algorithm 1 is initially and recursively feasible for every $\Phi(x^0) \in B$, i.e. at time $\hat{t} = t^0$ and at each successor time $\hat{t} \in t^0 + \delta\mathbb{N}$ the OCP (6) has a solution. In particular, the closed-loop system consisting of (1) and the FMPC feedback (7) has a global solution $x : [t^0, \infty) \to \mathbb{R}^n$ and the corresponding input is given by

$$u_{\mathrm{FMPC}}(t) = \mu(t, x(\hat{t})), \quad t \in [\hat{t}, \hat{t}+\delta),\ \hat{t} \in t^0 + \delta\mathbb{N}_0.$$

Furthermore, each global solution x with corresponding input u_{FMPC} satisfies:

(i) $\forall\, t \geq t^0 :\ |u_{\mathrm{FMPC}}(t)| \leq M$.

(ii) $\forall\, t \geq t^0 :\ \big|e^{(i)}(t)\big| < \psi_i(t)$ for $i = 0, 1$; in particular the error $e = y - y_{\mathrm{ref}}$ evolves within the funnel $\mathcal{F}_{\psi_0}$ and $\dot{e}$ within $\mathcal{F}_{\psi_1}$.

4. SIMULATION

To demonstrate the application of the FMPC Algorithm 1, we consider the example of a mass-spring system mounted on a car from Seifried and Blajer (2013). The mass m_2 moves on a ramp inclined by the angle $\vartheta \in [0, \frac{\pi}{2})$ and mounted on a car with mass m_1 by a spring-damper system, see Figure 2. It is possible to control the force $F = u$ acting on the car. The system is described by the equations

$$\begin{bmatrix} m_1 + m_2 & m_2\cos(\vartheta) \\ m_2\cos(\vartheta) & m_2 \end{bmatrix} \begin{pmatrix} \ddot{z}(t) \\ \ddot{s}(t) \end{pmatrix} + \begin{pmatrix} 0 \\ ks(t) + d\dot{s}(t) \end{pmatrix} = \begin{pmatrix} u(t) \\ 0 \end{pmatrix}, \tag{8}$$

where $z(t)$ is the horizontal position of the car and $s(t)$ the relative position of the mass on the ramp at time t. The physical constants $k > 0$ and $d > 0$ are the coefficients of the spring and damper, resp. The horizontal position of the mass on the ramp is the output y of the system, i.e.

$$y(t) = z(t) + s(t)\cos(\vartheta). \tag{9}$$

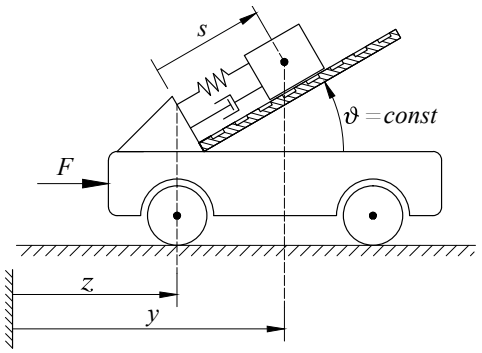


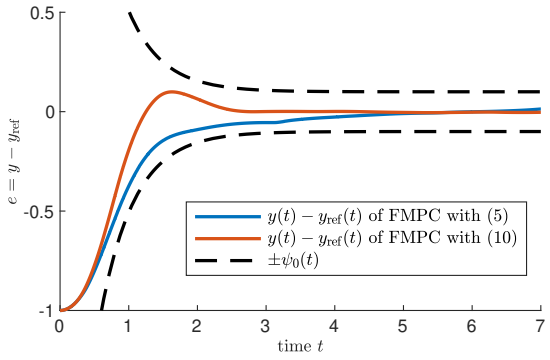
Fig. 2. Mass-on-car system.

We choose the parameters $m_1 = 4$, $m_2 = 1$, $k = 2$, $d = 1$, $\vartheta = \frac{\pi}{4}$ and initial values $z(0) = s(0) = \dot{z}(0) = \dot{s}(0) = 0$ for the simulation. The objective is tracking of the reference signal $y_{\mathrm{ref}} : t \mapsto \cos(t)$ so that the error $t \mapsto e(t) := y(t) - y_{\mathrm{ref}}(t)$ satisfies $|e(t)| \leq \psi_0(t)$ and $|\dot{e}(t)| \leq \psi_1(t)$ for all $t \geq 0$ with $\psi_0(t) = 3\,\mathrm{e}^{-2t} + 0.1$ and $\psi_1(t) = 6e^{-t} + 0.1$. One can easily verify that $\psi = (\psi_0, \psi_1) \in \mathcal{G}^1$ and that the initial errors lie within their respective funnel boundaries. The system (8) with output (9) has relative degree $r = 2$ for the given parameters. We compare the FMPC Algorithm 1 with stage cost (5) to the FMPC scheme from Berger et al. (2021) which uses the stage cost function $\tilde{\ell} : \mathbb{R}_{\geq 0} \times \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ with

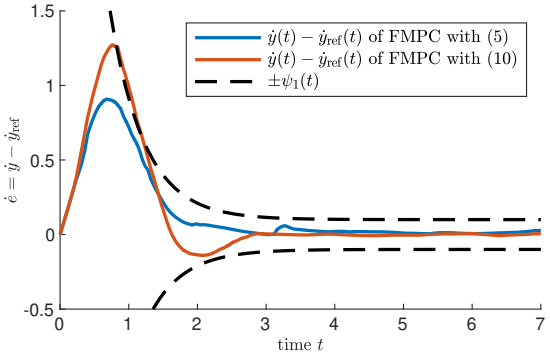
$$\tilde{\ell}(t, \zeta, u) = \begin{cases} \frac{1}{1 - \|e_0(t,\zeta)\|^2 / \psi_0(t)^2} + \lambda_u \|u\|^2, & \|e_0(t, \zeta)\| \neq \psi_0(t) \\ \infty, & \text{else.} \end{cases} \tag{10}$$

Contrary to the stage cost function (5), the function $\tilde{\ell}$ penalizes only the distance of the tracking error $e(t) = y(t) - y_{\mathrm{ref}}(t)$ to the funnel boundary ψ_0 but not derivative $\dot{e}(t) = \dot{y}(t) - \dot{y}_{\mathrm{ref}}(t)$ to the boundary ψ_1. For the FMPC Algorithm 1 the prediction horizon $T = 0.6$ and time shift $\delta = 0.04$. Due to discretisation, only step functions with constant step length 0.04 are considered for the OCP (6). We further choose for both stage cost functions the parameter $\lambda_u = 5 \cdot 10^{-3}$ and allow a maximal control value of $M = 30$. All simulations are performed on the time interval $[0, 7]$ with the MATLAB routine `ode45` and are depicted in Figure 3. Figure 3a shows the tracking error of the two different FMPC schemes evolving within the funnel boundaries given by ψ_0, while Figure 3b displays the derivative of the error within the boundaries given by ψ_1 The respective control signals generated by the controllers is displayed in Figure 3c.

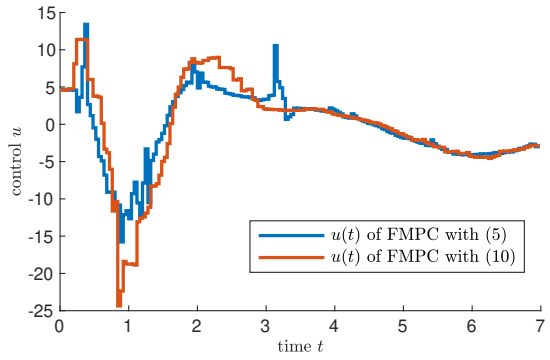
It is evident that both control schemes achieve the tracking of the reference signal within the performance boundaries given by ψ_0. While the FMPC Algorithm 1 with stage cost function (5) also ensures that the derivative of the tracking error evolves within funnel given by ψ_1, FMPC scheme with stage cost function $\tilde{\ell}$ as in (10) fails to do that and thus does not achieve the overall control objective. This is not surprising since the function $\tilde{\ell}$ does not penalize the distance of error's derivative to the funnel boundary.



(a) Funnel given by ψ_0 and tracking error e



(b) Funnel given by ψ_1 and tracking error derivative $\dot{e}$



(c) Control input

Fig. 3. Simulation of system (8) with output (9) under FMPC Algorithm 1 and FMPC from Berger et al. (2021)

Moreover, the FMPC Algorithm 1 with stage cost function (5) exhibits a smaller range of employed control values as the FMPC scheme from Berger et al. (2021).

5. CONCLUSION

In this note we outline a conceptual framework to extend the FMPC scheme proposed in Berger et al. (2021), which solves the problem of tracking a reference signal within a prescribed performance funnel, to systems with relative degree two. By exploiting concepts from funnel control and using a "funnel-like" stage cost, feasibility is achieved without the need for additional terminal or explicit output constraints while also being restricted to (a priori) bounded control values. In particular, additional output constraints in the OCP of FMPC as considered in Berger et al. (2020) and Berger and Dennstädt (2022) are not required to infer the feasibility results. However, contrary to previous results the prediction horizon has to be sufficiently long in order to guarantee recursive feasibility of the Funnel MPC algorithm. Extending these results to multi-input multi-output systems and systems with arbitrary relative degree $r > 2$ is subject of future work.

REFERENCES

Berger, T. and Dennstädt, D. (2022). Funnel MPC with feasibility constraints for nonlinear systems with arbitrary relative degree. *IEEE Control Systems Letters*, 6, 2804–2809.

Berger, T., Dennstädt, D., Ilchmann, A., and Worthmann, K. (2021). Funnel MPC for nonlinear system with relative degree one. Submitted for publication. Preprint available on arXiv: `https://arxiv.org/abs/2107.03284`.

Berger, T., Kästner, C., and Worthmann, K. (2020). Learning-based Funnel-MPC for output-constrained nonlinear systems. *IFAC-PapersOnLine*, 53(2), 5177–5182.

Byrnes, C.I. and Isidori, A. (1991). Asymptotic stabilization of minimum phase nonlinear systems. *IEEE Trans. Autom. Control*, 36(10), 1122–1137.

Chen, W.H., O'Reilly, J., and Ballance, D.J. (2003). On the terminal region of model predictive control for nonlinear systems with input/state constraints. *International journal of adaptive control and signal processing*, 17(3), 195–207.

Coron, J.M., Grüne, L., and Worthmann, K. (2020). Model predictive control, cost controllability, and homogeneity. *SIAM Journal on Control and Optimization*, 58(5), 2979–2996.

González, A.H. and Odloak, D. (2009). Enlarging the domain of attraction of stable MPC controllers, maintaining the output performance. *Automatica*, 45(4), 1080–1085.

Grüne, L. and Pannek, J. (2017). *Nonlinear Model Predictive Control: Theory and Algorithms*. Springer, London.

Hackl, C.M., Hopfe, N., Ilchmann, A., Mueller, M., and Trenn, S. (2013). Funnel control for systems with relative degree two. *SIAM J. Control Optim.*, 51(2), 965–995.

Ilchmann, A., Ryan, E.P., and Sangwin, C.J. (2002). Tracking with prescribed transient behaviour. *ESAIM: Control, Optimisation and Calculus of Variations*, 7, 471–493.

Manrique, T., Fiacchini, M., Chambrion, T., and Millérioux, G. (2014). MPC tracking under time-varying polytopic constraints for real-time applications. In *2014 European Control Conference (ECC)*, 1480–1485. IEEE.

Qin, S.J. and Badgwell, T.A. (2003). A survey of industrial model predictive control technology. *Control engineering practice*, 11(7), 733–764.

Rawlings, J.B., Mayne, D.Q., and Diehl, M. (2017). *Model predictive control: theory, computation, and design*, volume 2. Nob Hill Publishing Madison, WI.

Seifried, R. and Blajer, W. (2013). Analysis of servo-constraint problems for underactuated multibody systems. *Mechanical Sciences*, 4, 113–129.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Separation of Nonlinearity and Stochasticity in Nonlinear Diffusion Control Problems ⋆

William M. McEneaney * **Peter M. Dower** ** **Tao Wang** ***

* *Dept. Mech. and Aero. Eng., UC San Diego, La Jolla, CA 92093 USA, (e-mail: wmceneaney@ucsd.edu)*
** *Dept. Elec. & Electronic Eng., Univ. Melbourne, Victoria 3010, Australia, (e-mail: pdower@unimelb.edu.au)*
*** *Dept. Mech. and Aero. Eng., UC San Diego, La Jolla, CA 92093 USA, (e-mail: taw003@eng.ucsd.edu)*

Abstract: A class of nonlinear, stochastic staticization control problems (including minimization problems with smooth, convex, coercive payoffs) driven by diffusion dynamics with constant diffusion coefficient is considered. The nonlinearities are addressed through staticization-based duality. The second-order Hamilton-Jacobi partial differential equations (HJ PDEs) are converted into associated control problems with higher-dimensional states. In these problems, one component of the state propagates by deterministic, nonlinear dynamics, while the other component is a scaled Brownian motion. These components interact only through a bilinear terminal cost. This structure will be exploited to generate an efficient solution approach.

Keywords: dynamic programming, Hamilton-Jacobi, partial differential equations, stochastic control, staticization.

1. INTRODUCTION

The results herein can be equivalently viewed as results regarding Hamilton-Jacobi partial differential equations (HJ PDEs), or as results regarding the optimal control of systems with dynamics defined by stochastic differential equations (SDEs). The second-order Hamilton-Jacobi partial differential equations (HJ PDEs) are converted into associated control problems with higher-dimensional states. In contrast to the original problems, in these associated problems, one component of the state propagates by deterministic, nonlinear dynamics, while the other component is a scaled Brownian motion. These components interact only through a bilinear terminal cost. This structure will be exploited to generate an efficient solution approach. It will be shown that numerical solutions of many nonlinear second-order HJ PDE problems may be obtained through numerical solutions of nonlinear first-order HJ PDE problems, along with associated finite-dimensional sets of differential Riccati equations (DREs). In more general cases, solution of steady-state second-order HJB PDEs in potentially much lower dimension may be required.

Consider the HJ PDE

$$0 = W_t + \underset{v\in\mathbb{R}^k}{\text{stat}} \left\{ f^T(x,v)W_x + L(x,v) \right\} + \tfrac{1}{2}\,\text{tr}[AW_{xx}], \quad (1)$$

for $(t,x) \in (0,T)\times\mathbb{R}^n$, where specific assumptions on the problem data appear further below. The "stat" operator is briefly discussed in Section 2.2, but it is useful to note here that in the case of a convex, coercive, C^1 argument, it is equivalent to the minimization operator, and hence the results herein typically subsume HJ PDEs of the form

$$0 = W_t + \min_{v\in\mathbb{R}^k} \left\{ f^T(x,v)W_x + L(x,v) \right\} + \tfrac{1}{2}\,\text{tr}[AW_{xx}].$$

Consider the following class of quadratic terminal costs. Let $\mathcal{M}_{2n}$ denote the set of matrices $\bar{\Pi}$ such that

$$\bar{\Pi} \doteq \begin{pmatrix} \bar{M} & -\bar{M} \\ -\bar{M} & \bar{M} \end{pmatrix},$$

where $\bar{M}$ is symmetric and nonsingular, and let

$$\begin{aligned} W(T,x) &= \psi(x; z, \bar{\pi}, \bar{\Pi}, \bar{\gamma}) \\ &\doteq \tfrac{1}{2} \begin{pmatrix} x \\ z \end{pmatrix}^T \bar{\Pi} \begin{pmatrix} x \\ z \end{pmatrix} + \bar{\pi}^T \begin{pmatrix} x \\ z \end{pmatrix} + \bar{\gamma}, \end{aligned} \quad (2)$$

where $\bar{\Pi} \in \mathcal{M}_{2n}$, $z \in \mathbb{R}^n$, $\bar{\pi} \in \mathbb{R}^{2n}$, $\bar{\gamma} \in \mathbb{R}$ are parameters.

It is well-known that for purposes of numerical solution, nonlinear stochastic control problems are typically converted into the second-order HJ PDE problems such as (1) with associated terminal data, and further, the dimension of the space over which these PDEs are defined is that of the state process of the control problem. Of course, realistic control problems typically have relatively high dimensional state processes, leading to PDEs over high dimensional spaces. The solution of such HJ PDE problems has long been hampered by the curse-of-dimensionality, which refers to the fact that with classical algorithms, the computational cost grows exponentially fast as a function of space dimension, and we note that this has limited the solvability of such problems by classical methods to state-space dimensions on the order of three to five, c.f. Falcone and Ferretti (2014). More recently, max-plus based curse-of-dimensionality-free methods (in addition to other notable new approaches) have demonstrated computational tractability for certain classes of problems in significantly

⋆ Research partially supported by AFOSR and NSF.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

higher space dimension, and this approach have been quite effective in the case of first-order HJ PDEs [Akian, Gaubert and Lakhoua (2008); Dower (2018); Gaubert et al (2011); McEneaney (2006); Qu (2014)].

2. PRELIMINARIES

2.1 Problem Class and Recollection of Results

We consider a nonlinear stochastic control problem with SDE dynamics and initial state are given by

$$d\xi(t) = f(\xi(t), \tilde{v}(t))\, dt + \bar{a}\, dB(t), \quad \xi(s) = x \in \mathbb{R}^n, \quad (3)$$

where the underlying probability space is denoted as $(\Omega, \mathcal{F}_\infty, P)$ with Ω denoting the sample space, $\mathcal{F}_\infty$ denoting the σ-algebra and P denoting the probability measure. Also, $B(t)$ denotes an n-dimensional Brownian motion adapted to filtration $\mathcal{F}_t$. Assumptions on f will be indicated further below. We suppose the controls take values in $V = \mathbb{R}^k$. Fix $T \in (0, \infty)$, and for $s \in [0, T]$, let

$$\mathcal{V}_{s,T} \doteq \{\tilde{v} : [s, T] \times \Omega \to \mathbb{R}^k | \tilde{v} \text{ is } \mathcal{F}_\cdot\text{-adapted, right-contin. and such that } \mathbb{E}\textstyle\int_s^T |\tilde{v}(t)|^m\, dt < \infty \ \forall m \in I\!N\},$$

$$\|\tilde{v}\| = \|\tilde{v}\|_{\mathcal{V}_{s,T}} \doteq \max_{m \le \breve{M}} \Big[\mathbb{E} \int_s^T |\tilde{v}(t)|^m\, dt\Big]^{1/m},$$

where $\breve{M} \doteq 8\breve{q}$ and $\breve{q}$ will be specified in Assumption ($\mathcal{A}$.1). The payoff will be given by

$$J(s, x, \tilde{v}; z, \bar{\Pi}, \bar{\pi}, \bar{\gamma}, T) \doteq \mathbb{E}\Big\{ \int_s^T L(\xi(t), \tilde{v}(t))\, dt + \psi(\xi(T); z)\Big\}$$

where ψ is as indicated in (2), $\bar{\pi} \in \mathbb{R}^{2n}$, $\bar{\gamma} \in \mathbb{R}$ and $z \in \mathbb{R}^n$. In the more general case, one takes a terminal cost form

$$\Psi(x) = \Psi(x; z) \doteq \operatorname*{stat}_{z \in \mathbb{R}^n} \Big\{ \tfrac{1}{2} \begin{pmatrix} x \\ z \end{pmatrix}^T \bar{\Pi} \begin{pmatrix} x \\ z \end{pmatrix} + a(z) \Big\}, \quad (4)$$

with some specified function, $a(\cdot)$, where the definition of operator stat follows in Section 2.2. The terminal cost in (4) is a "stat-quad" representation [Dower and McEneaney (2020); McEneaney and Dower (2018)], of a general class of terminal costs that may be represented as such. We will consider only the terminal cost form (2) here. For $(s, x) \in [0, T] \times \mathbb{R}^n$, the value function is given by

$$\bar{W}(s, x; z, \bar{\Pi}, \bar{\pi}, \bar{\gamma}, T) \doteq \operatorname*{stat}_{\tilde{v} \in \mathcal{V}_{s,T}} J(s, x, \tilde{v}; z, \bar{\Pi}, \bar{\pi}, \bar{\gamma}, T), \quad (5)$$

where $\bar{\pi} = 0$ here, but will be generalized below.

We will proceed through several steps. The first step is to obtain the equivalence between the value function and the solution of the associated HJ PDE problem. This equivalence is standard in the optimization and game cases, and less so in staticization cases that are not equivalent to these. We recall a result in the staticization case where the stationary value is given by (3)–(5). Specifically, letting $\mathcal{X}_0 \doteq (0, T) \times \mathbb{R}^n$, $\bar{\mathcal{X}}_0 \doteq (0, T] \times \mathbb{R}^n$, $z \in \mathbb{R}^n$, $\bar{\pi} \in \mathbb{R}^{2n}$, $\bar{\gamma} \in \mathbb{R}$ and $\bar{\Pi} \in \mathcal{M}_{2n}$, consider

$$0 = W_t + \operatorname*{stat}_{v \in \mathbb{R}^k} \big\{ f(x, v)^T W_x + L(x, v) \big\} + \tfrac{1}{2} \operatorname{tr}[A W_{xx}] \quad (6)$$

$$\doteq W_t + H_0(x, W_x) + \mathcal{Q}_0(x, W_x) + \tfrac{1}{2} \operatorname{tr}[A W_{xx}], \quad (7)$$

$$(s, x) \in \mathcal{X}_0,$$

$$W(T, x; z, \bar{\Pi}, \bar{\gamma}) = \psi(x; z, \bar{\pi}, \bar{\Pi}, \bar{\gamma}), \quad x \in \mathbb{R}^n, \quad (8)$$

where $A = \bar{a}\bar{a}^T$, $\mathcal{Q}_0$ is a quadratic function of its arguments, and the non-quadratic components of the Hamiltonian are isolated within H_0.

We make the following assumptions.

Assume that for $z \in \mathbb{R}^n$, there exists $W = W(\cdot, \cdot; z) \in C^{1,4}(\mathcal{X}_0) \cap C_p(\bar{\mathcal{X}}_0)$ satisfying (7)–(8) (where C_p denotes the space of continuous functions with at most quadratic growth), and that there exists $\bar{C}_0 < \infty$ and $\breve{q} \in I\!N$ such that $|W_x(s, x)| \le \bar{C}_0(1 + |x|^{2\breve{q}})$ and $|W_{xx}(s, x)| \le \bar{C}_0(1 + |x|^{2\breve{q}})$ for all $(s, x) \in \bar{\mathcal{X}}_0$. Assume $\bar{M}$ is positive definite, symmetric; $f, L \in C^3(\mathbb{R}^n \times \mathbb{R}^k)$; $\exists \bar{C}_1 < \infty$ such that for all $x, v \in \mathbb{R}^n$ $|f_x(x, v)|, |f_v(x, v)| \le \bar{C}_1$, $|f_{xx}(x, v)|, |f_{xv}(x, v)|, |f_{vv}(x, v)| \le \bar{C}_1$ and $|L_{xx}(x, v)|, |L_{xv}(x, v)|, |L_{vv}(x, v)|, \le \bar{C}_1$. Assume that for each $z \in \mathbb{R}^n$, there exists $v^* \in C(\bar{\mathcal{X}}_0)$ that is globally Lipschitz in x, and is such that $v^*(t, x) \in \operatorname{argstat}_{v \in \mathbb{R}^k}\{f(x, v)^T W_x(t, x) + L(x, v)\}$ for all $(t, x) \in \mathcal{X}_0$. ($\mathcal{A}$.1)

Theorem 1. Assume ($\mathcal{A}$.1). Then $W = \bar{W}$ on $\bar{\mathcal{Z}}$, and v^* is a stationary control yielding payoff $\bar{W}$. Further, with $\bar{\xi}^*$ denoting the trajectory generated by v^*, and letting $\tilde{v}^*(t) \doteq v^*(t, \bar{\xi}^*(t))$ for all t, $\tilde{v}^*$ is continuous with respect to t.

The proof of Theorem 1 may be found in McEneaney, Dower and Wang (2021). All results to follow are obtained under ($\mathcal{A}$.1).

2.2 Staticization and Stat-Quad Duality

Let $\mathcal{V}$ denote a Banach space over either the real or complex field, and suppose $\mathcal{U}$ is a [real or complex] normed vector space with $\mathcal{A} \subseteq \mathcal{U}$, and suppose $G : \mathcal{A} \to \mathcal{F}$. Let $\operatorname{argstat}_{u \in \mathcal{A}} G(u) \doteq \operatorname{argstat}\{G(u) \,|\, u \in \mathcal{A}\}$, $\operatorname{stat}^s{}_{u \in \mathcal{A}} G(u) \doteq \big\{G(\bar{u}) \,\big|\, \bar{u} \in \operatorname{argstat}\{G(u) \,|\, u \in \mathcal{A}\}\big\}$ and $\operatorname{stat}_{u \in \mathcal{A}} G(u) \doteq a$ be defined as in, for example, [McEneaney, Dower and Wang (2021); McEneaney and Zhao (2019); McEneaney and Dower (2018)].

Analogous to semiconvex duality, we have the following "stat-quad" duality (McEneaney and Dower, 2018, Th. 4). In the nondegenerate case, it is closely related to the Legendre-Fenchel transform. Let ϕ denote a generic function in $C^j(\mathbb{R}^n; \mathbb{R})$ with $j \ge 2$ such that there exists $\bar{c}_2 < \infty$ such that $|\phi_{uu}(u)| \le \bar{c}_2$ for all $u \in \mathbb{R}^n$. We let $\psi_0(u, w) \doteq \frac{\bar{m}}{2}|u - w|^2$ for all $u, w \in \mathbb{R}^n$, where $\bar{m} \in \mathbb{R} \setminus \{0\}$.

Lemma 2. Suppose $|\bar{m}| > \bar{c}_2$. We have

$$\phi(u) = \operatorname*{stat}_{w \in \mathbb{R}^n} [a(w) + \psi_0(u, w)] \quad \forall\, u \in \mathbb{R}^n, \quad (9)$$

$$a(w) = \operatorname*{stat}_{u \in \mathbb{R}^n} [\phi(u) - \psi_0(u, w)] \quad \forall\, w \in \mathbb{R}^n. \quad (10)$$

Further, there exists unique $u^* \in C^{j-1}(\mathbb{R}^n; \mathbb{R}^n)$ such that

$$a(w) = \phi(u^*(w)) - \psi_0(u^*(w), w), \quad \forall\, w \in \mathbb{R}^n, \quad (11)$$

$$a_w(w) = \bar{m}(u^*(w) - w) = \phi_u(u^*(w)) \quad \forall\, w \in \mathbb{R}^n, \quad (12)$$

where $a \in C^j(\mathbb{R}^n; \mathbb{R})$. Suppose $|\bar{m}| \in [2\bar{c}, \infty)$. Then, for each $u \in \mathbb{R}^n$, there exists unique argstat $w^*(u)$. Further, $w^* \in C^{j-1}(\mathbb{R}^n; \mathbb{R}^n)$, and one has

$$\phi(u) = a(w^*(u)) + \psi_0(u, w^*(u)), \quad w^* = (u^*)^{-1}.$$

Remark 3. For simplicity here, we are assuming that the condition $|\bar{m}| \in (\bar{c}_2, \infty)$ is satisfied, as it provides a simple sufficient condition. However, one should note that implies convexity or concavity, while stat-quad duality is employed so that we may cover a wider class of cases.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

3. A STATICIZATION-BASED REPRESENTATION OF THE NONLINEARITIES

For $x, p, \alpha, \beta \in \mathbb{R}^n$, let

$$\mathcal{Q}(x,p,\alpha,\beta) \doteq \tfrac{c_1}{2}|x-\alpha|^2 + \tfrac{c_2}{2}|p-\beta|^2, \tag{13}$$

where $c_1, c_2 \in \mathbb{R}$. We make the following assumption, which will be sufficient to guarantee existence of all the relevant duality objects to follow.

We assume that $H_0 \in C^{\hat{m}}(\mathbb{R}^{2n})$ with $\hat{m} \geq 4$, and that the second derivatives of H_0 are uniformly bounded. ($\mathcal{A}.2$)

As an aid in developing results further below, we recall the following from McEneaney, Dower and Wang (2021).

Lemma 4. Let $|c_1|, |c_2|$ be sufficiently large. Then,

$$H_0(x,p) = \operatorname*{stat}_{(\alpha,\beta)\in\mathbb{R}^{2n}} \big[G_0(\alpha,\beta) + \mathcal{Q}(x,p,\alpha,\beta)\big].$$

$$G_0(\alpha,\beta) = \operatorname*{stat}_{(x,p)\in\mathbb{R}^{2n}} \big[H_0(x,p) - \mathcal{Q}(x,p,\alpha,\beta)\big],$$

$\operatorname*{argstat}_{(x,p)\in\mathbb{R}^{2n}} \big[H_0(x,p) - \mathcal{Q}(x,p,\alpha,\beta)\big]$ is single-valued,

$\operatorname*{argstat}_{(\alpha,\beta)\in\mathbb{R}^{2n}} \big[G_0(\alpha,\beta) + \mathcal{Q}(x,p,\alpha,\beta)\big]$ is single-valued,

and $G_0 \in C^{\hat{m}}(\mathbb{R}^{2n})$ with bounded second derivatives.

Now, we let the coefficients in $\mathcal{Q}_0$ be specifically given by

$$\begin{aligned}\mathcal{Q}_0(x,p) &= \tfrac{1}{2}\begin{pmatrix}x\\p\end{pmatrix}^T D \begin{pmatrix}x\\p\end{pmatrix} + d^T\begin{pmatrix}x\\p\end{pmatrix}\\ &= \tfrac{1}{2}\begin{pmatrix}x\\p\end{pmatrix}^T\begin{pmatrix}D_{1,1} & D_{1,2}\\ D_{2,1} & D_{2,2}\end{pmatrix}\begin{pmatrix}x\\p\end{pmatrix} + \begin{pmatrix}d_1\\d_2\end{pmatrix}^T\begin{pmatrix}x\\p\end{pmatrix},\end{aligned}$$

where D is symmetric. Note that for $|c_2|$ sufficiently large,

$$\begin{aligned}&G_0(\alpha,\beta) + \mathcal{Q}_0(x,p) + \mathcal{Q}(x,p,\alpha,\beta) \\ &= \operatorname*{stat}_{v\in\mathbb{R}^n}\big\{[D_{1,2}x + d_2 - c_2\beta + v]^T p + H_1(x,\alpha,\beta,v)\big\},\end{aligned} \tag{14}$$

where

$$\begin{aligned}&H_1(x,\alpha,\beta,v) \doteq G_0(\alpha,\beta) + \tfrac{1}{2}x^T D_{1,1}x + \tfrac{c_1}{2}|x-\alpha|^2\\ &+ \tfrac{c_2}{2}|\beta|^2 + d_1^T x + \tfrac{1}{2}v^T\Gamma v,\\ &\Gamma \doteq -(c_2\mathcal{I}_n + D_{2,2})^{-1}.\end{aligned}$$

Consider the following stationarity control problem. Let the dynamics be given by

$$\begin{aligned}d\xi(t) &= f'(\xi(t),\bar{\beta}(t),\tilde{v}(t))\,dt + \bar{a}\,dB(t)\\ &\doteq (D_{1,2}\xi(t) + d_2 - c_2\bar{\beta}(t) + \tilde{v}(t))\,dt + \bar{a}\,dB(t)\end{aligned} \tag{15}$$

where $\tilde{v}, \bar{\beta} \in \mathcal{V}_{s,T}$. Let the payoff and stationary value be

$$\begin{aligned}J'(s,x,\tilde{v},\bar{\alpha},\bar{\beta};z,T) &\doteq \mathbb{E}\Big\{\int_s^T H_1(\xi(t),\bar{\alpha}^*(t),\bar{\beta}(t),\tilde{v}(t))\,dt\\ &\quad + \psi(\xi(T);z)\Big\},\\ W'(s,x;z,T) &\doteq \operatorname*{stat}_{(\tilde{v},\bar{\alpha},\bar{\beta})\in\mathcal{V}_{s,T}\times[\mathcal{O}_{s,T}]^2} J'(s,x,\tilde{v},\bar{\alpha},\bar{\beta};z,T),\end{aligned} \tag{16}$$

$$\begin{aligned}\mathcal{O}_{s,T} \doteq \{&\nu : [s,T]\times\Omega \to \mathbb{R}^k \,|\, \nu \text{ is } \mathcal{F}.\text{-adapted, contin.}\\ &\text{w.r.t. time, and s.t. } \mathbb{E}\textstyle\int_s^T |\nu(t)|^2\,dt < \infty\,\}.\end{aligned}$$

Also consider the iterated form of W' given by

$$\hat{W}'(s,x;z,T) \doteq \operatorname*{stat}_{(\bar{\alpha},\bar{\beta})\in[\mathcal{O}_{s,T}]^2} \operatorname*{stat}_{\tilde{v}\in\mathcal{V}_{s,T}} J'(s,x,\tilde{v},\bar{\alpha},\bar{\beta};z,T). \tag{17}$$

Theorem 5. Let $|c_1|, |c_2|$ be sufficiently large. For each $z \in \mathbb{R}^n$, value function W' is identical to the value function, $\bar{W}$ (given in (5)). Further, there exists unique (α^*, β^*, v^*) such that

$$\begin{aligned}[\alpha^*,\beta^*,v^*](t,x) \in \operatorname*{argstat}_{(\alpha,\beta,v)\in\mathbb{R}^{3n}} &\big\{[f'(x,\beta,v)]^T\bar{W}_x(t,x)\\ &+ H_1(x,\alpha,\beta,v)\big\},\end{aligned}$$

and $[\alpha^*,\beta^*,v^*](t,\xi(t))$ is a staticizing control. Lastly, for each $z \in \mathbb{R}^n$, value function $\hat{W}'$ is identical to value function $\bar{W}$.

4. SEPARATION

We now change the point of view to that of a control problem with state variables being both π and x. We will see that although in this case the dynamics are not purely deterministic, the nonlinear deterministic component and the linear stochastic component are separated in a very useful way, where in particular, the only interaction is through a bilinear terminal cost term. First, note that the inner staticization of (17) is a set of linear-quadratic Gaussian control problems, indexed by the $\bar{\alpha}, \bar{\beta}$. Consider the dynamics

$$\begin{aligned}\dot{\hat{\Pi}}(t) &= -\bar{F}_1(\hat{\Pi}(t))\\ &\doteq -\big\{\hat{\Pi}(t)K_2\hat{\Pi}(t) + K_3^T\hat{\Pi}(t) + \hat{\Pi}(t)K_3 + K_1\big\},\end{aligned} \tag{18}$$

$$\begin{aligned}\dot{\hat{\pi}}(t) &= -\bar{F}_2(\hat{\Pi}(t),\hat{\pi}(t),\bar{\alpha}(t),\bar{\beta}(t))\\ &\doteq -\big\{\hat{\Pi}(t)K_2\hat{\pi}(t) + \hat{\Pi}(t)\hat{\mathcal{I}}^{1,1}V^2(t) + K_3\hat{\pi}(t) + V^1(t)\big\},\end{aligned} \tag{19}$$

$$\begin{aligned}\dot{\hat{\gamma}}(t) &= -\bar{F}_3(\hat{\Pi}(t),\hat{\pi}(t),\bar{\alpha}(t),\bar{\beta}(t))\\ &\doteq -\big\{G_0(\bar{\alpha}(t),\bar{\beta}(t)) + \tfrac{c_1}{2}|\bar{\alpha}(t)|^2 + \tfrac{c_2}{2}|\bar{\beta}(t)|^2\\ &\quad + \tfrac{1}{2}\hat{\pi}(t)^T K_2\hat{\pi}(t) + (V^2(t))^T\hat{\pi}(t) + \tfrac{1}{2}\operatorname{tr}(K_4\hat{\Pi}(t)K_5)\big\},\end{aligned} \tag{20}$$

with terminal conditions, following from (2), given by

$$\hat{\Pi}(T) = \bar{\Pi} \doteq \begin{pmatrix}\bar{M} & -\bar{M}\\ -\bar{M} & \bar{M}\end{pmatrix},$$

$\hat{\pi}(T) = \bar{\pi} \doteq 0$, $\hat{\gamma}(T) = \bar{\gamma}$, where

$$K_1 \doteq \begin{pmatrix}k_1 & 0\\ 0 & 0\end{pmatrix},\ K_2 \doteq \begin{pmatrix}k_2 & 0\\ 0 & 0\end{pmatrix},\ K_3 \doteq \begin{pmatrix}D_{1,2}^T & 0\\ 0 & 0\end{pmatrix},$$

$$K_4 \doteq (A\ 0),\ K_5 \doteq \begin{pmatrix}\mathcal{I}_n\\ 0\end{pmatrix},\ \hat{\mathcal{I}}^{1,1} \doteq \begin{pmatrix}\mathcal{I}_n & 0\\ 0 & 0\end{pmatrix},$$

$$V^1(t) \doteq \begin{pmatrix}d_1 - c_1\bar{\alpha}(t)\\ 0\end{pmatrix},\ V^2(t) \doteq \begin{pmatrix}d_2 - c_2\bar{\beta}(t)\\ 0\end{pmatrix},$$

$k_1 \doteq c_1\mathcal{I}_n + D_{1,1}$ and $k_2 \doteq c_2\mathcal{I}_n + D_{2,2}$.

Let $(\hat{\Omega}, \hat{\mathcal{F}}_\infty, \hat{P})$ denote a probability space, and let $\hat{B}$ with $\hat{B}(\hat{s}) = 0$ a.s., be a Brownian motion on $(\hat{\Omega}, \hat{\mathcal{F}}_\infty, \hat{P})$, adapted to filtration $\hat{\mathcal{F}}.$. Finally, let

$$\begin{aligned}\hat{\mathcal{V}}_{\hat{s},\hat{t}} \doteq \big\{&\hat{v} : [\hat{s},\infty)\times\hat{\Omega} \to \mathbb{R}^k | \hat{v} \text{ is } \hat{\mathcal{F}}.\text{-adapted, rt.-contin.,}\\ &\text{s.t. } \hat{v}(r) = 0 \text{ a.s. } \forall r > \hat{t},\ \mathbb{E}\textstyle\int_{\hat{s}}^{\hat{t}} |\hat{v}(t)|^m\,dt < \infty\ \forall m \in I\!N\big\}.\end{aligned} \tag{21}$$

For $\hat{s} \in [0,\hat{t})$, consider the dynamics

$$\dot{\pi}(r) = \bar{F}_2(\Pi(r),\pi(r),\bar{\hat{\alpha}}(r),\bar{\hat{\beta}}(r)), \quad \pi(\hat{s}) = \bar{\pi}, \tag{22}$$

$$\dot{\gamma}(r) = \bar{F}_3(\Pi(r),\pi(r),\bar{\hat{\alpha}}(r),\bar{\hat{\beta}}(r)), \quad \gamma(\hat{s}) = \bar{\gamma}, \tag{23}$$

$$\xi(r) = x + \bar{a}\hat{B}(r), \tag{24}$$

where $(\bar{\hat{\alpha}}, \bar{\hat{\beta}}) \in \hat{\mathcal{V}}^2_{\hat{s},\hat{t}}$. We also let $\bar{\bar{\mu}} \doteq (\bar{\hat{\alpha}}, \bar{\hat{\beta}})$. Define the decomposition of π into $\rho, q \in \mathbb{R}^n$, by $(\rho^T, q^T) \doteq \pi^T$. Letting $y \doteq (x^T, z^T)^T$, consider

$$\hat{J}(\hat{s}, (\bar{\pi}, x), \bar{\bar{\mu}}; z, \bar{\Pi}, \hat{t}) \doteq \mathbb{E}\Big\{ \pi^T(\hat{t}) \begin{pmatrix} \xi(\hat{t}) \\ z \end{pmatrix} + \gamma(\hat{t}) \Big\}$$

$$= \mathbb{E}\Big\{ y^T \pi(\hat{t}) + \big(\bar{a}\hat{B}(\hat{t})\big)^T \rho(\hat{t}) + \int_{\hat{s}}^{\hat{t}} \bar{F}_3(\Pi(r), \pi(r), \bar{\bar{\mu}}(r))\, dr \Big\} + \bar{\gamma}, \tag{25}$$

$$\hat{W}(\hat{s}, (\bar{\pi}, x); z, \bar{\Pi}, \hat{t}) \doteq \operatorname*{stat}_{\bar{\bar{\mu}} \in \hat{\mathcal{V}}^2_{\hat{s},\hat{t}}} \hat{J}(\hat{s}, (\bar{\pi}, x), \bar{\bar{\mu}}; z, \bar{\Pi}, \hat{t}). \tag{26}$$

One should note that the only *direct* interaction between the nonlinear dynamics of π and the diffusion process is through the bilinear term in the terminal cost.

In order to map the results of Theorem 1 to the the case of HJ PDE problem (27)–(28) below and control problem (22)–(26), we make the following definitions. Let

$$\bar{\bar{x}} \doteq \begin{pmatrix} \bar{\pi} \\ x \end{pmatrix} = \begin{pmatrix} \bar{\rho} \\ \bar{q} \\ x \end{pmatrix}, \quad \bar{\bar{\xi}}(\cdot) \doteq \begin{pmatrix} \pi(\cdot) \\ \xi(\cdot) \end{pmatrix} = \begin{pmatrix} \rho(\cdot) \\ q(\cdot) \\ \xi(\cdot) \end{pmatrix},$$

$$\bar{\bar{f}}(\bar{\Pi}, \bar{\bar{\xi}}, \bar{\bar{\mu}}) \doteq \begin{pmatrix} \bar{F}_2(\bar{\Pi}, \pi, \bar{\bar{\mu}}) \\ 0 \end{pmatrix}, \quad \bar{\bar{A}} \doteq (0_{n\times 2n}, \bar{a}\mathcal{I}_n)^T,$$

$$\bar{\bar{L}}(\bar{\Pi}, \bar{\bar{\xi}}, \bar{\bar{\mu}}) \doteq \bar{F}_3(\bar{\Pi}, \pi, \bar{\bar{\mu}}), \quad \bar{\bar{\psi}}(\bar{\bar{\xi}}; z, \bar{\Pi}, \bar{\gamma}) \doteq \psi(\xi; z, \bar{\Pi}, \bar{\gamma}).$$

We implicitly invoke trivial congruences such as that between $(0, \hat{t}) \times \mathbb{R}^{2n} \times \mathbb{R}^n$ and $(0, \hat{t}) \times \mathbb{R}^{3n}$ without further mention, and let $\mathcal{X}^2_{0,\hat{t}} \doteq (0, \hat{t}) \times \mathbb{R}^{2n}$, $\bar{\mathcal{X}}^2_{0,\hat{t}} \doteq [0, \hat{t}] \times \mathbb{R}^{2n}$, $\mathcal{X}^3_{0,\hat{t}} \doteq (0, \hat{t}) \times \mathbb{R}^{3n}$ and $\bar{\mathcal{X}}^3_{0,\hat{t}} \doteq [0, \hat{t}] \times \mathbb{R}^{3n}$.

We also consider the HJ PDE problem

$$0 = W_{\hat{s}} + \operatorname*{stat}_{\bar{\bar{\mu}} \in \mathbb{R}^{2n}} \big\{ W^T_{\bar{\bar{x}}} \bar{\bar{f}}(\Pi(\hat{t} - \hat{s}), \bar{\bar{x}}, \bar{\bar{\mu}}) + \bar{\bar{L}}(\Pi(\hat{t} - \hat{s}), \bar{\bar{x}}, \bar{\bar{\mu}}) \big\} + \tfrac{1}{2} \operatorname{tr}[AW_{xx}], \quad (\hat{s}, \bar{\bar{x}}) \in (0, \hat{t}) \times \mathbb{R}^{3n}, \tag{27}$$

$$W(\hat{t}, \bar{\bar{x}}; \bar{\gamma}) = y^T \pi + \bar{\gamma}, \quad \bar{\bar{x}} \in \mathbb{R}^{2n}, \tag{28}$$

where Π satisfies $\dot{\Pi} = \bar{F}_1(\Pi)$. We make the following assumptions, which are similar to their analogues in ($\mathcal{A}$.1).

($\mathcal{A}$.3) Fix $z \in \mathbb{R}^n$, $\bar{\Pi} \in \mathbb{R}^{2n\times 2n}$, $\bar{\gamma} \in \mathbb{R}$ and $\hat{t} \in (0, T)$. Assume there exists a solution, $(\hat{s}, \bar{\pi}, x) \mapsto W(\hat{s}, \bar{\pi}, x; z, \bar{\Pi}, \bar{\gamma}, \hat{t})$, of (27)–(28) in $C^{1,4}(\mathcal{X}^3_{0,\hat{t}}) \cap C_p(\bar{\mathcal{X}}^3_{0,\hat{t}})$, such that there exist $\bar{c}_3, \bar{C}_0 < \infty$ and $\hat{q} \in I\!N$ such that $|W_{xxxx}(\hat{s}, \bar{\bar{x}})| \le \bar{c}_3$, $|W_{\bar{\bar{x}}}(\hat{s}, \bar{\bar{x}})| \le \bar{C}_0(1 + |\bar{\bar{x}}|^{2\hat{q}})$ and $|W_{\bar{\bar{x}}\bar{\bar{x}}}(\hat{s}, \bar{\bar{x}})| \le \bar{C}_0(1 + |\bar{\bar{x}}|^{2\hat{q}})$, for all $(\hat{s}, \bar{\bar{x}}) \in \bar{\mathcal{X}}^3_{0,\hat{t}}$. Assume $\bar{\bar{f}}, \bar{\bar{L}} \in C^3(\mathbb{R}^{2n\times 2n} \times \mathbb{R}^{3n} \times \mathbb{R}^{2n})$; $\exists \bar{C}_1 < \infty$ such that for all $(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}}) \in \mathbb{R}^{2n\times 2n} \times \mathbb{R}^{3n} \times \mathbb{R}^{2n}$, $|\bar{\bar{f}}_{\bar{\bar{x}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})|, |\bar{\bar{f}}_{\bar{\bar{\mu}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})| \le \bar{C}_1$, $|\bar{\bar{f}}_{\bar{\bar{x}}\bar{\bar{x}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})|, |\bar{\bar{f}}_{\bar{\bar{x}}\bar{\bar{\mu}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})|, |\bar{\bar{f}}_{\bar{\bar{\mu}}\bar{\bar{\mu}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})| \le \bar{C}_1$ and $|\bar{\bar{L}}_{\bar{\bar{x}}\bar{\bar{x}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})|, |\bar{\bar{L}}_{\bar{\bar{x}}\bar{\bar{\mu}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})|, |\bar{\bar{L}}_{\bar{\bar{\mu}}\bar{\bar{\mu}}}(\bar{\Pi}, \bar{\bar{x}}, \bar{\bar{\mu}})|, \le \bar{C}_1$. There exists $\bar{\mu}^* \in C(\bar{\mathcal{X}}^3_{0,\hat{t}})$ that is globally Lipschitz in $\bar{\bar{x}}$, and is such that $\bar{\mu}^*(r, \bar{\bar{x}}) \in \operatorname{argstat}_{\bar{\mu} \in \mathbb{R}^{2n}} \{ \bar{\bar{f}}(\Pi(r), \bar{\bar{x}}, \bar{\mu})^T W_{\bar{\bar{x}}}(r, \bar{\bar{x}}) + \bar{\bar{L}}(\Pi(r), \bar{\bar{x}}, \bar{\mu}) \}$ for all $(r, \bar{\bar{x}}) \in \bar{\mathcal{X}}^3_{0,\hat{t}}$.

Theorem 6. $W = \hat{W}$ on $[\hat{s}, \hat{t}] \times \mathbb{R}^{3n}$, and $\bar{\mu}^*$ is a staticizing feedback control yielding payoff $\hat{W}$. Further, with $\hat{v}^*(r) \doteq \bar{\mu}^*(r, \bar{\xi}^*(r))$ for all $r \in [\hat{s}, \hat{t}]$, $\hat{v}^*$ is a.s. continuous.

Fix $z \in \mathbb{R}^n$, $\bar{\Pi} \in \mathbb{R}^{2n\times 2n}$, $\bar{\gamma} \in \mathbb{R}$ and $\hat{t} \in (0, T)$, and let

$$\hat{W}^f(\hat{t}, (\pi, x); z, \bar{\Pi}, \bar{\gamma}) \doteq \bar{G}(0, x, z, \hat{t}; \bar{\Pi}) + \hat{W}(0, (\pi, x); z, \bar{\Pi}, \bar{\gamma}, \hat{t}) \quad \forall\, (\hat{t}, (\pi, x)) \in \bar{\mathcal{X}}^3_{0,T}.$$

Theorem 7. Let $T \in (0, \infty)$, $z \in \mathbb{R}^n$, $\bar{\Pi} \in \mathbb{R}^{2n\times 2n}$, $\bar{\gamma} \in \mathbb{R}$. Then, for all $(s, (\pi, x)) \in \bar{\mathcal{X}}^3_{0,T}$,

$$\hat{W}^f(T - s, (\pi, x); z, \bar{\Pi}, \bar{\gamma}) = \bar{W}(s, x; z, \bar{\Pi}, \pi, \bar{\gamma}, T),$$

where the latter is defined in (5).

Remark 8. Note that HJ PDE problem (27)–(28) is equivalently

$$0 = W_{\hat{s}} + \tfrac{1}{2} \operatorname{tr}[AW_{xx}] + \operatorname*{stat}_{\bar{\mu} \in \mathbb{R}^{2n}} \big\{ W^T_\pi \bar{F}_2(\Pi(\hat{t} - \hat{s}), \pi, \bar{\mu}) + \bar{F}_3(\Pi(\hat{t} - \hat{s}), \pi, \bar{\mu}) \big\}, \quad (\hat{s}, \pi, x) \in (0, \hat{t}) \times \mathbb{R}^{3n}, \tag{29}$$

$$W(\hat{t}, \pi, x; z, \bar{\gamma}) = y^T \pi + \bar{\gamma}, \quad (\pi, x) \in \mathbb{R}^{3n}. \tag{30}$$

Note that x does not appear inside the stat operation in (29). Alternatively, note that the stochastic component of control problem (22)–(26) interacts with the nonlinear component of the problem only through a bilinear expression in the terminal cost. This structure is greatly simplified over that of the original problem, and this will be exploited in the generation of efficient solutions.

REFERENCES

M. Akian, S. Gaubert, and A. Lakhoua, *The max-plus finite element method for solving deterministic optimal control problems: Basic properties and convergence analysis,* SIAM J. Control and Optim., 47 (2008), 817–848.

P.M. Dower and W.M. McEneaney, "Verification of stationary action trajectories via optimal control", Proc. 2020 Amer. Control Conf. (Denver), 1779–1784.

P.M. Dower, "An adaptive max-plus eigenvector method for continuous time optimal control problems", *Numerical methods for optimal control,* Eds. M. Falcone and R. Ferretti and L. Grune and W.M. McEneaney, Springer, INDAM Series (2018), 211–240.

M. Falcone and R. Ferretti, *Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations*, SIAM, 2014.

S. Gaubert, W.M. McEneaney and Z. Qu, "Curse of dimensionality reduction in max-plus based approximation methods: theoretical estimates and improved pruning algorithms", Proc. IEEE CDC/ECC 2011, 1054–1061.

W.M. McEneaney, P.M. Dower and T. Wang, "Second-Order Hamilton-Jacobi PDE Problems and Certain Related First-Order Problems, Part 1: Approximation", SIAM J Control and Optim., (submitted, 2021).

W.M. McEneaney and R. Zhao, "Staticization and Iterated Staticization," SIAM J Control and Optim., to appear.

W.M. McEneaney and P.M. Dower, "Static duality and a stationary-action application", J. Diffl. Eqs., 264 (2018), 525–549.

W.M. McEneaney and H. Kaise, "Idempotent expansions for continuous-time stochastic control", SIAM J. Control and Optim., 54 (2016), 73–98.

W.M. McEneaney, *Max-Plus Methods for Nonlinear Control and Estimation*, Birkhauser, Boston, 2006.

Z. Qu, "A max-plus based randomized algorithm for solving a class of HJB PDEs", Proc. IEEE CDC, Dec. 2014.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On discrete-time optimal control and the related bounded real and positive real lemmas

Edward Branford * **Timothy H. Hughes** *

* *University of Exeter, Penyrn Campus*
(e-mail: E.H.Branford@exeter.ac.uk, T.H.Hughes@exeter.ac.uk).

Abstract: We extend the classical discrete-time bounded real lemma to the general case of systems that need not be controllable or observable, and its relationship to the related discrete-time optimal control problem. In the talk accompanying this extended abstract, we will further discuss the analogies in discrete time to the recent continuous time development of an assumption free theory of linear passive and non-expansive systems that draws on the behavioral framework of Jan Willems and collaborators.

Keywords: Dissipativity, Optimal Control, Discrete-time linear systems

1. INTRODUCTION

We present a new version of the discrete-time bounded real lemma (see Theorem 4). This new version addresses outstanding assumptions in existing literature, as discussed following that theorem statement.

1.1 Notation

A small amount of notation is in order for this work. Firstly, we let $\mathbb{R}$ denote the real numbers. We let $\mathbb{R}[z]$ denote the polynomials in the indeterminate z, and $\mathbb{R}(z)$ denote the rational functions in the indeterminate z. We denote matrices with entries in the real numbers as $\mathbb{R}^{m\times n}$, and equivalently for matrices with polynomial entries ($\mathbb{R}^{m\times n}[z]$) and entries in the rational functions ($\mathbb{R}^{m\times n}(z)$). If M is a real- or complex-valued matrix, then M^T denotes its transpose and M^* denotes its Hermitian transpose, while for a matrix of rational functions $M(z)$, $M^\sim(z) = (M(\frac{1}{z}))^T$. Finally, a matrix $M \in \mathbb{R}^{m\times m}$ is called non-negative definite if it is symmetric (i.e., $M = M^T$) and $\mathbf{z}^T M\mathbf{z} \geq 0$ for all real vectors $\mathbf{z} \in \mathbb{R}^m$, and positive definite if, in addition, $\mathbf{z}^T M\mathbf{z} = 0$ implies $\mathbf{z} = 0$.

1.2 Background

This work concerns discrete-time dissipativity of linear time-invariant systems, defined as follows:

Definition 1. (Discrete-time dissipative system). We consider the discrete-time state-space system

$$\Sigma \coloneqq \{\mathbf{x}(k{+}1) = A\mathbf{x}(k) + B\mathbf{u}(k) \text{ and } \mathbf{y}(k) = C\mathbf{x}(k) + D\mathbf{u}(k), \text{ for } k = 0, 1, 2, \dots \};$$

where A, B, C, D are appropriately dimensioned real-valued matrices, and we let d denote the dimension of the state $\mathbf{x}$. We refer to the solutions $(\mathbf{u}, \mathbf{y}, \mathbf{x})$ of this equation as trajectories of the system. We define the supply rate

$$s(\mathbf{u}, \mathbf{y}) \coloneqq \mathbf{u}^T\mathbf{u} - \mathbf{y}^T\mathbf{y},$$

and we say that the system Σ is dissipative with respect to the supply rate s if there exists a function $M : \mathbb{R}^d \to \mathbb{R}$ such that

$$-\sum_{i=0}^{N} s(\mathbf{u}, \mathbf{y})(i) \leq M(\mathbf{x}_0)$$

for all $N \geq 0$ and all trajectories $(\mathbf{u}, \mathbf{y}, \mathbf{x})$ of the system Σ with $\mathbf{x}(0) = \mathbf{x}_0$. We then define the available storage $S_a(\mathbf{x}_0)$ as the least such function M, i.e.,

$$S_a(\mathbf{x}_0) \coloneqq \sup_{(\mathbf{u},\mathbf{y},\mathbf{x})\in\Sigma, N\geq 0, \mathbf{x}(0)=\mathbf{x}_0} -\sum_{i=0}^{N} s(\mathbf{u}, \mathbf{y})(i).$$

This represents the discrete-time analogy to the definition of continuous-time dissipativity of Willems (1972a,b). The definition extends in a natural way to other supply functions. Such dissipative systems as outlined in the above definition correspond to the systems under consideration in the famous discrete-time bounded real lemma. The purpose of this extended abstract is to present a new version of the discrete-time bounded real lemma that completely characterises the conditions under which a system satisfies the dissipativity condition as defined above, in the absence of any assumptions whatsoever.

Firstly, we define a discrete bounded real function (hereafter denoted DBR):

Definition 2. (DBR). Let $G \in \mathbb{R}^{m\times n}(z)$. G is DBR if (i) G is analytic throughout the exterior of the unit circle; and (ii) $I - G(z)^*G(z) \geq 0$ for all complex numbers z whose magnitude is strictly greater than one.

Secondly, we define the observability matrix for the system Σ:

$$\mathcal{O} \coloneqq \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix}, \tag{1}$$

whose columns are linearly independent if and only if the system Σ is observable in the usual sense of the word.

We also recall a recent result on discrete-time spectral factors that will be useful in characterising the available storage. This result, from Baggio and Ferrante (2016), provides a relatively recent discrete-time version of the long-standing result in continuous-time by Youla (1961).

Lemma 3. Let $H \in \mathbb{R}^{n\times n}(z)$ and let $\max_{\lambda\in\mathbb{C}}(\text{rank}(H(\lambda)))$ $= r$. Further, suppose that $H(\lambda)$ is non-negative definite for all λ on the unit circle, with the exception of poles of H. Then there exists a rational matrix $Z \in \mathbb{R}^{r\times n}(z)$ such that (i) $H = Z^{\sim}Z$; (ii) Z is analytic throughout the exterior of the unit circle; (iii) the rows of $Z(\lambda)$ are independent for all λ outside of the unit circle; and (iv) $\lim_{\lambda\to\infty} Z(\lambda)$ exists and its rows are independent. If Z satisfies conditions (i)—(iv), then we call Z a discrete-time spectral factor of H, and if H has no poles on the unit circle then Z has no poles on the unit circle.

2. THEOREM STATEMENT

The new version of the discrete-time bounded real lemma, whose proof is the subject of a journal paper in preparation, is then as follows:

Theorem 4. Let Σ be as in Definition 1 and let $\mathcal{O}$ be as in equation (1). The following statements are equivalent

(1) Σ is dissipative with respect to the supply rate $s = \mathbf{u}^T\mathbf{u} - \mathbf{y}^T\mathbf{y}$, in accordance with Definition 1.
(2) Let $G(z) = D + C(zI - A)^{-1}B$ and let $U \in \mathbb{R}^{n\times n}[z]$ and $V \in \mathbb{R}^{n\times d}[z]$ be left coprime polynomial matrices that satisfy $U(z)B^T\mathcal{O}^T = V(z)(\frac{1}{z}I - A^T)\mathcal{O}^T$ (such matrices will always exist). The following three conditions all hold:
   (a) G is DBR, in accordance with Definition 2
   (b) If $\mathbf{z}$ is a complex-valued vector, λ is a complex number whose modulus is greater than or equal to one, and $\mathbf{z}^T\mathcal{O}[\lambda I - A \quad B] = 0$, then $\mathbf{z}^T\mathcal{O} = 0$.
   (c) If $\mathbf{b} \in \mathbb{R}^n[z]$ satisfies $\mathbf{b}^T(UU^{\sim} - (VC^T + UD^T)(VC^T + UD^T)^{\sim}) = 0$, then there exists a polynomial vector $X \in \mathbb{R}^d[z]$ such that $(\mathbf{b}^T(VC^T + UD^T))(z)C = X(z)^T(zI - A)$;
(3) There exists a real-valued non-negative definite matrix P such that the block matrix
$$\begin{bmatrix} P - A^TPA - C^TC & -C^TD - A^TPB \\ -D^TC - B^TPA & I - D^TD - B^TPB \end{bmatrix} \tag{2}$$
is non-negative definite.
(4) There exists a real-valued non-negative matrix P_-, and real-valued matrices L and W such that the following three conditions all hold:
   (a) $\begin{bmatrix} P_- - A^TP_-A - C^TC & -C^TD - A^TP_-B \\ -D^TC - B^TP_-A & I - D^TD - B^TP_-B \end{bmatrix}$
   $= \begin{bmatrix} L^T \\ W^T \end{bmatrix}[L \; W]$;
   (b) with the notation $Z(z) = W + L(zI - A)^{-1}B$, then Z is a discrete-time spectral factor of $I - G^{\sim}G$ in accordance with Definition 3;
   (c) if $\mathbf{z}$ is a real-valued vector satisfying $\mathcal{O}\mathbf{z} = 0$, then $P_-\mathbf{z} = 0$.

Moreover, if the above conditions hold, then $S_a(\mathbf{x}_0) = \mathbf{x}_0^TP_-\mathbf{x}_0$ for all $\mathbf{x}_0 \in \mathbb{R}^d$, where P_- is as in condition 4, and if P is a non-negative definite matrix satisfying condition 3, then $P \geq P_-$.

Together these conditions form a complete characterisation of the dissipativity of a general linear discrete-time system with respect to the supply rate s, the existence of non-negative definite solutions to the matrix inequality in condition 3, a characterisation of the spectral factorisation of the function $I - G^{\sim}G$, and a characterisation of the available storage whenever the system is dissipative. The equivalence of conditions 2–4 are analogous to results in the context of discrete-time positive-real systems by Branford and Hughes (2020). That paper did not consider the connection to dissipativity and the associated optimal control problem, and we note that a similar connection can also be made in the setting of discrete-time positive-real systems relating to dissipativity with respect to the supply rate $s_p(\mathbf{u}, \mathbf{y}) = \mathbf{u}^T\mathbf{y}$. Specifically, conditions (1)–(3) in Branford and Hughes (2020, Theorem 5) are equivalent to the system in Definition 1 being dissipative with respect to the supply rate s_p, and the available storage in this case takes the form $S_a(\mathbf{x}_0) = \frac{1}{2}\left(\mathbf{x}_0^T\tilde{P}_-\mathbf{x}_0\right)$ where (i) $\tilde{P}_-$ is as in Branford and Hughes (2020, Theorem 5, condition 3); and (ii) if $\mathbf{z}$ is a real-valued vector satisfying $\mathcal{O}\mathbf{z} = 0$, then $\tilde{P}_-\mathbf{z} = 0$.

The key distinction of the preceding theorem from existing results in the literature is the absence of any a-priori restrictions on the eigenvalues of the A matrix or the controllability or observability of the system. The consequence in terms of the second set of conditions in the theorem statement is that the familiar condition that $G(z)$ must be a discrete-time bounded real function (see, e.g., Vaidyanathan, 1985) must be augmented with two further conditions.

The first of these additional conditions, condition 2b, is equivalent to the requirement that the system be behaviorally stabilizable. In other words, for any given real k_0, any past input-output trajectory $(\mathbf{u}(k), \mathbf{k}(k))$, $k < k_0$, can be concatenated with a future trajectory $(\tilde{\mathbf{u}}(k), \tilde{\mathbf{y}}(k))$, $k \geq k_0$ with the property that $\tilde{\mathbf{u}}(k) \to 0$ and $\tilde{\mathbf{y}}(k) \to 0$ as $k \to \infty$. Note that this is not the same as state stabilizability, e.g., if D satisfies $I - D^TD \geq 0$ and $B = C = 0$ then the system is in fact dissipative for any given matrix A (indeed, $P = 0$ then satisfies condition 2 of Theorem 4).

Condition 2c is a coupling condition between lossless trajectories of the system (i.e., trajectories for which, were they returned to the same state at $k = N$ as their initial state at $k = 0$, would satisfy $\sum_{i=0}^{N} s(\mathbf{u}, \mathbf{y})(i) = 0$), and trajectories for which $\mathbf{u}(k) = 0$ for all k. This condition can be tested by first computing the left syzygy of $UU^{\sim} - (VC^T + UD^T)(VC^T + UD^T)^{\sim}$ (i.e., a polynomial matrix $H(z)$ with as many rows as the minimum dimension of the left kernel of $(UU^{\sim} - (VC^T + UD^T)(VC^T + UD^T)^{\sim})(\lambda)$ over all complex values of λ, and for which the rows of $H(\lambda)$ are always in this left kernel and are independent for all complex values of λ). Such a polynomial matrix can be obtained using standard symbolic algebra software. We note here that $UU^{\sim} - (VC^T + UD^T)(VC^T + UD^T)^{\sim}$ may be the ratio of a polynomial matrix and the monomial z^p for some integer p, whereby in order to compute the left syzygy using standard techniques for polynomial matrices it may be necessary to multiply this matrix by the monomial z^p to obtain a polynomial matrix whose left syzygy will be identical to that of $UU^{\sim} - (VC^T +$

$UD^T)(VC^T + UD^T)^\sim$. Next, compute the polynomial matrix $H(VC^T+UD^T)C$, which we shall denote K. Then, expand $K(z)$ in terms of z: $K(z) = K_0+K_1z+K_2z^2+\cdots$. Condition 2c is then equivalent to requiring $K_0 + K_1A + K_2A^2 + \cdots = 0$.

It is appropriate to recognise other notable contributions to generalise the applicability of the discrete-time bounded real lemma. For example, the paper by de Souza and Xie (1992), in which the assumption of controllability is removed, but the a-priori assumption is made that the eigenvalues of A are all strictly within the unit circle and, moreover, that $I - D^TD - B^TPB > 0$. In addition, Xiao and Hill (1999) present a necessary and sufficient condition for a transfer function $G(z) = D + C(zI - A)^{-1}B$ to be DBR in the absence of controllability and observability assumptions that involves the matrix in (2) and the controllability matrix

$$\mathcal{C} = \begin{bmatrix} B & AB & \cdots & A^{d-1}B \end{bmatrix}. \tag{3}$$

A simple example serves to illustrate the improvement in the version of the discrete-time bounded real lemma presented here. Consider a system Σ with $A = 0, B = 0, C = 1$ and $D = 1$. Then $D+C(zI-A)^{-1}B = 1$ which is a discrete bounded real function. But $I-D^TD-B^TPB = 0$, so if $I - D^TD - B^TPB = W^TW$ then we require $W = 0$. But this implies that $-1 = -C^TD - A^TPB = L^TW = 0$, a contradiction. Note also that in this case there does not exist a $P \geq 0$ such that $I - D^TD - B^TPB$ is non-singular. Hence, this case is not considered in the version of the discrete-time bounded real lemma presented in de Souza and Xie (1992). Moreover, the system satisfies the necessary and sufficient conditions outlined in Xiao and Hill (1999, Lemma 8) involving the matrix in (2) and the controllability matrix in equation (3), yet the system is not dissipative. It can further be verified using the method outlined in the previous paragraph that condition 2c does not hold for this example. In contrast, it can be shown that the system Σ with $A = 0, B = 1, C = 0$ and $D = 1$ does satisfy condition 2 and is in fact dissipative.

Finally, we note that Definition 1 and the results presented thereafter are specific to systems for which the difference equation stated in that definition is only required to hold for non-negative k. This implies that the initial state $\mathbf{x}(0)$ can be arbitratily specified. For the system Σ considered in the previous paragraph, in the case that $A = 0, B = 0, C = 1$ and $D = 1$, then if the difference equation in Definition 1 is instead required to hold for all k including negative values it follows that $\mathbf{y}(k) = \mathbf{u}(k)$ for all k, whereupon $-\sum_{i=0}^{N} s(\mathbf{u}, \mathbf{y})(i)$ is bounded above by zero. In contrast, when the difference equation only holds for non-negative k, then $-\sum_{i=0}^{N} s(\mathbf{u}, \mathbf{y})(i)$ can be made arbitrarily large for any given $\mathbf{x}(0) \neq 0$ by picking an initial input $\mathbf{u}(0)$ with the same sign as $\mathbf{x}(0)$ and a sufficiently large magnitude.

REFERENCES

Baggio, G. and Ferrante, A. (2016). On the factorization of rational discrete-time spectral densities. *IEEE Transactions on Automatic Control*, 61(4), 969–981.

Branford, E. and Hughes, T. (2020). An assumption-free theorem on discrete-time positive real systems. *IFAC-PapersOnLine*, 53(2), 4474–4480.

de Souza, C.E. and Xie, L. (1992). On the discrete-time bounded real lemma with application in the characterization of static state feedback $\|H\|_\infty$ controllers. *Systems and Control Letters*, 18, 61–71.

Vaidyanathan, P.P. (1985). The discrete-time bounded-real lemma in digital filtering. *IEEE Trans. on Circuits and Systems*, CAS-32(9), 918–924.

Willems, J.C. (1972a). Dissipative dynamical systems, Part I: General theory. *Arch. Ration. Mech. Anal.*, 45, 321–351.

Willems, J.C. (1972b). Dissipative dynamical systems, Part II: Linear systems with quadratic supply rates. *Arch. Ration. Mech. Anal.*, 45, 352–393.

Xiao, C. and Hill, D.J. (1999). Generalizations and new proof of the discrete-time positive real lemma and bounded real lemma. *IEEE Transactions on Circuits and Systems*, 46(6).

Youla, D.C. (1961). On the factorization of rational matrices. *IRE Transactions on Information Theory*, 7.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Simulating Mid-Air Interaction Trajectories via Model Predictive Control

Markus Klar * **Florian Fischer** * **Arthur Fleig** *
Miroslav Bachinski * **Jörg Müller** *

* *University of Bayreuth, 95440 Bayreuth, Germany*
(e-mail: markus.klar@uni-bayreuth.de)

Abstract: We investigate the ability of Model Predictive Control (MPC) to generate human-like movements during interaction with mid-air user interfaces, i.e., pointing in virtual or augmented reality, using a state-of-the-art biomechanical model. The model is partly a black box implemented in the MuJoCo physics engine, requiring either gradient-free optimization algorithms or gradient approximation. This makes it even more important to choose the objective function or the MPC horizon length wisely. We introduce three objective functions suggested in the literature and identify optimal cost weights such that the simulated trajectories best match real ones obtained from motion capturing, i.e., we tackle an inverse optimal control problem. For the best performing objective function, we then analyze the effects of the horizon length and of the cost weights. This model-based approach enables the analysis of interaction techniques, e.g., in terms of ergonomics and effort, without the need for extensive user studies.

Keywords: Human Computer Interaction, Modeling of human performance, Model Predictive Control, Work in real and virtual environments
AMS subject classifications: 68U20, 68U07, 90C90

1. INTRODUCTION

Recent achievements in virtual and augmented reality have brought up a variety of new interaction methods, ranging from typing in mid-air to complex manipulation of virtual objects. Naturally, such interaction techniques include a number of parameters (offsets, rotations, gains, etc.) that have a significant impact on the resulting interaction experience. As a result, careful tuning of these parameters is critical. Typically, this fine-tuning requires multiple user studies and therefore takes a considerable amount of time. These user studies could be partially replaced by simulations, where movement is calculated by solving Optimal Control Problems (OCPs).

This optimal control approach differs from traditional approaches in the field of Human Computer Interaction (HCI), which are rather evidence-based and focus on summary statistics such as Fitts' Law (Fitts, 1954) and phenomena like bell-shaped velocity profiles and corrective submovements. In contrast, we make use of recent models and methods from the fields of Control Theory and Optimal Feedback Control (OFC), which allows us to analyze the continuous development of any quantity of interest (e.g., end-effector position or muscle activations) (Müller et al., 2017). Building on such a control-theoretic forward simulation of interactive movement, we are able to identify optimal weights of a suitable objective function, i.e., solve an inverse optimal control problem (Albrecht, 2013). We thus have two layers of optimization: The inner layer calculates optimal movement trajectories w.r.t. a given objective function. The outer layer computes optimal weights of the objective function such that the simulated trajectories best match real ones obtained from motion capture. Combining this rigorous mathematical framework with an efficient implementation of a state-of-the-art biomechanical model has the potential to simulate and predict human mid-air trajectories during interaction – as we show in this work by applying Model Predictive Control (MPC) to a model of mid-air pointing.

2. MODELING

We consider a discrete-time MPC framework [1], i.e., we consider the following optimal control problem

$$\begin{aligned} &\min_{\mathbf{u}(\cdot)\in\mathbb{U}^N} \sum_{n=0}^{N-1} \ell(\mathbf{x}(n),\mathbf{u}(n)) \\ \text{s.t.}\quad &\mathbf{x}(n+1) = f(\mathbf{x}(n),\mathbf{u}(n)) \quad \forall n \in \{0,...,N-1\}, \\ &\mathbf{x}(0) = \mathring{\mathbf{x}} \end{aligned} \tag{1}$$

on a receding horizon of length $N \in \mathbb{N}$, only applying the first part of the optimal control sequence, $\mathbf{u}^*(0)$, in every MPC step. In the remainder of this section, we explain and motivate all variables appearing in (1).

2.1 System Dynamics

The components of the system dynamics f in (1) are illustrated in Figure 1. In addition to the "Interaction Technique" block, f encodes the human upper extremity body dynamics, which is split in the two blocks "Muscle Model" and "MuJoCo".

To simulate movement of the human body we use the fast physics simulation software MuJoCo (Todorov et al.,

[1] For more details on nonlinear MPC see Grüne and Pannek (2017).

2012). In our previous work (Fischer et al., 2021), we have derived a MuJoCo model from a state-of-the-art OpenSim (Seth et al., 2018) model of the upper extremity (Saul et al., 2014). This musculoskeletal model includes a shoulder, an elbow, and a wrist, with 7 independent joints, i.e., degrees of freedom (DOFs) (3 in the shoulder, 2 in the elbow, 2 in the wrist), and 13 coupled joints. To generate arm movements, MuJoCo applies torque at each independent joint. Since humans are not capable of producing torque instantaneously, we incorporate the following second-order muscle model.

Given a normalized control signal $\mathbf{u} \in \mathbb{U} := [-1,1]^7$ from the MPC block in Figure 1, we model basic muscle properties such as activation delays for each independent joint $i = 1, ..., 7$ via the second-order dynamics

$$\begin{bmatrix} \sigma_i(k+1) \\ \Delta\sigma_i(k+1) \end{bmatrix} = A \cdot \begin{bmatrix} \sigma_i(k) \\ \Delta\sigma_i(k) \end{bmatrix} + B \cdot u_i, \tag{2a}$$

$$\sigma_i(0) = \mathring{\sigma}_i, \quad \Delta\sigma_i(0) = \Delta\mathring{\sigma}_i, \quad k \in \{0, ..., K-1\}, \tag{2b}$$

with

$$A = \begin{bmatrix} 1 & \Delta t \\ \frac{-\Delta t}{(t_e t_a)} & 1 - \Delta t \frac{t_e + t_a}{t_e t_a} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \frac{\Delta t}{t_e t_a} \end{bmatrix}, \tag{3}$$

which yields the activation vector $\boldsymbol{\sigma} \in [-1,1]^7$ and its approximate derivative $\Delta\boldsymbol{\sigma} \in \mathbb{R}^7$. Here, $K \in \mathbb{N}$ denotes the number of steps, $\Delta t = 2$ ms is the sampling time, and $t_e = 30$ ms and $t_a = 40$ ms are the excitation and activation time constants, respectively, as in Van der Helm and Rozendaal (2000).

The biomechanical simulation in MuJoCo (cf. Figure 1) takes these activations $\boldsymbol{\sigma}$ as input and yields new joint angles $\mathbf{q} \in \mathbb{R}^7$ and velocities $\mathbf{v} \in \mathbb{R}^7$ as well as the position of the fingertip $\mathbf{y} \in \mathbb{R}^3$. The subsequent Interaction Technique block maps the fingertip $\mathbf{y}$ to the virtual cursor position $\mathbf{p} \in \mathbb{R}^3$. The mapping can be a simple translation by an offset $\boldsymbol{\omega} \in \mathbb{R}^3$, i.e., $\mathbf{p} = \mathbf{y} + \boldsymbol{\omega}$, as in the "Virtual Cursor" case that we consider, cf. Section 3. More generally, it could represent a dynamic system with its own internal state, e.g., for pointer acceleration techniques (Müller, 2017).

In summary, the state vector $\mathbf{x}$ consists of the joint angles $\mathbf{q}$ and velocities $\mathbf{v}$, the activations $\boldsymbol{\sigma}$ and their approximate derivatives $\Delta\boldsymbol{\sigma}$, and the cursor position $\mathbf{p}$:

$$\mathbf{x} = (\mathbf{q}, \mathbf{v}, \boldsymbol{\sigma}, \Delta\boldsymbol{\sigma}, \mathbf{p})^\top. \tag{4}$$

The initial state is denoted by $\mathring{\mathbf{x}}$.

As humans are known to change control behavior not continuously but only intermittently (Gawthrop et al., 2011), the control signal $\mathbf{u}$ is assumed to only change every 40 ms, while both MuJoCo and the Muscle Model update every 2 ms. Thus, evaluating $f(\mathbf{x}(n), \mathbf{u}(n))$ in (1) involves $K = 20$ forward steps of the muscle model (2) and the MuJoCo simulation starting from $\mathbf{x}(n)$, in which the control signal $\mathbf{u} = \mathbf{u}(n)$ is kept constant.

2.2 Objective Functions

In the following, we introduce three different stage costs ℓ that were proposed in HCI and movement science to explain human behavior (Berret et al., 2011). Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm.

Distance and Control Cost (DC) A first approach is to penalize the remaining distance from the cursor to the target and adding quadratic control costs as regularization, i.e.,

$$\ell(\mathbf{x}, \mathbf{u}) = \|\mathbf{p} - \bar{\mathbf{p}}\| + r_1 \|\mathbf{u}\|^2, \tag{5}$$

where $\bar{\mathbf{p}} \in \mathbb{R}^3$ is the target center and $r_1 \geq 0$ can be interpreted as a weight for the trade-off between accuracy and effort. This was used in several works studying human movement, cf. Diedrichsen et al. (2010).

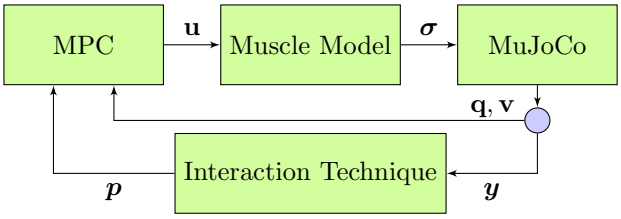


Fig. 1. High-level view on the closed feedback loop.

Joint Acceleration (JAC) The second approach is a simple extension of (5), where in addition high joint accelerations $\mathbf{a} = \dot{\mathbf{v}}$ are penalized:

$$\ell(\mathbf{x}, \mathbf{u}) = \|\mathbf{p} - \bar{\mathbf{p}}\| + r_1 \|\mathbf{u}\|^2 + r_2 \|\mathbf{a}\|^2, \tag{6}$$

where $r_1, r_2 \geq 0$ are weights. Since we cannot obtain joint accelerations $\mathbf{a}$ directly from MuJoCo, we use central differences of the velocities $\mathbf{v}$ (one-sided differences on the boundaries) to approximate them. From a biomechanical perspective, high joint acceleration leads to greater wear of the joints and should thus be avoided. Using inverse control, it could be shown that this cost term plays an important role in human movements (Berret et al., 2011).

Commanded Torque Change (CTC) As a third variant, we consider the commanded torque change cost from Nakano et al. (1999). In our case, CTC is directly proportional to $\Delta\boldsymbol{\sigma}$. Therefore, the stage cost reads

$$\ell(\mathbf{x}, \mathbf{u}) = \|\mathbf{p} - \bar{\mathbf{p}}\| + r_1 \|\mathbf{u}\|^2 + r_2 \|\Delta\boldsymbol{\sigma}\|^2, \tag{7}$$

where $r_1, r_2 \geq 0$ are weight parameters.

2.3 Numerical Solution

To solve OCP (1), we use direct single shooting to obtain a nonlinear program, which is solved with L-BFGS-B implemented in `scipy` [2]. Since gradient information is not available in closed-form due to partly blackbox dynamics, we approximate the gradients via central finite differences.

3. USER STUDY

We are interested in how well our model, with one of the stage costs (5)-(7), can synthesize human movement given different interaction techniques. To this end, we conducted a small user study with 6 participants (Mean Age=28.8, SD=6.6, 4 Male, all right-handed) from our local university campus. Participants wore a head mounted display [3] and a full-body suit to track their movements at 240Hz [4]. An LED marker was placed at the fingertip of the index finger to generate the virtual cursor position. Our experimental design involved two factors: interaction technique (Virtual Cursor vs. Virtual Pad) and specific

[2] https://scipy.org/
[3] HTC Vive Pro, https://vive.com/de/product/vive-pro/
[4] Phasespace X2E, https://phasespace.com/x2e-motion-capture/

setting (identity vs. ergonomic). Both factors were varied within subjects. The task was based on the discrete Fitts' Law paradigm, following the ISO 9241-9 standard: 13 virtual targets with a diameter of 5 cm were placed on a circle of 30 cm diameter, 50 cm in front of the right shoulder of the participant. Participants were instructed to move the cursor as fast as possible, only using their shoulder and arm. Each participant performed the complete ISO task consisting of 13 different targets 5 times per condition. After some preprocessing, where reaction times were removed as our model cannot account for them, we further removed 158 trials due to irregularities such as early starts or exceptionally long trials, leaving a total of 1402 movements, which we use in the following evaluation.

4. MODEL EVALUATION

In this section, we perform a pairwise comparison between simulated cursor trajectories to real user data. The simulated cursor trajectories, which we denote by $\boldsymbol{p}_{\text{sim}}$, are obtained by iteratively solving the OCP (1) on a receding horizon. This procedure is continued until the movement duration matches that of the user.

4.1 Quantitative Comparison of Cost Functions

To evaluate how well the simulated movement matches empirically observed user behavior, for each stage cost, we first perform a parameter optimization of r_1, r_2 for each user and condition separately, as follows. We choose five randomly selected trials per user and condition, and for each trial calculate the root mean squared error (RMSE)

$$\text{RMSE}(\boldsymbol{p}_{\text{sim}}, \boldsymbol{p}_{\text{real}})^2 = \frac{1}{M} \sum_{n=0}^{M-1} \|\boldsymbol{p}_{\text{sim}}(n) - \boldsymbol{p}_{\text{real}}(n)\|^2 . \quad (8)$$

Here, $\boldsymbol{p}_{\text{real}}(n) \in \mathbb{R}^3$ is the cursor position of the user at time step $n \in \mathbb{N}$, and $M \in \mathbb{N}$ is the number of time steps of the considered trial. We then use the *CMAES* (Hansen, 2016) to find cost weights that minimize the total RMSE between simulation and user position trajectories.[5] The identified cost weights are then used in the pairwise comparison of all 1402 movement trajectories.

To illustrate variability in human movement, we include both signal-dependent and constant control noise in the MPC closed-loop trajectory. That is, in every MPC step, after solving the OCP (1) (with no noise involved), we replace $\mathbf{u}^*(0)$ by a normally distributed random variable $\alpha(\mathbf{u}^*(0)) \sim \mathcal{N}(\mathbf{u}^*(0),\, 0.103^2 \cdot \mathbf{u}^*(0)^2 + 0.185^2)$, as suggested by van Beers et al. (2004). This illustrates a planned movement that, upon execution, is perturbed.

A box plot containing the RMSEs of all trials for each considered cost function is shown in Figure 2 (percentage of outliers: DC: 4.09%; CTC: 8.03%; JAC: 5.88%). Adding an additional cost term to the DC costs (5), as we have done for both the JAC (6) and the CTC costs (7), results in a significant (****, i.e., $p \leq 10^{-4}$) improvement of the RMSEs while also reducing variance. One has to keep in mind that both (6) and (7) introduce an additional weight parameter r_2, i.e., the observed better fit was to be

[5] To avoid the even more expensive gradient estimation at this point, we use the gradient-free evolutionary algorithm CMAES.

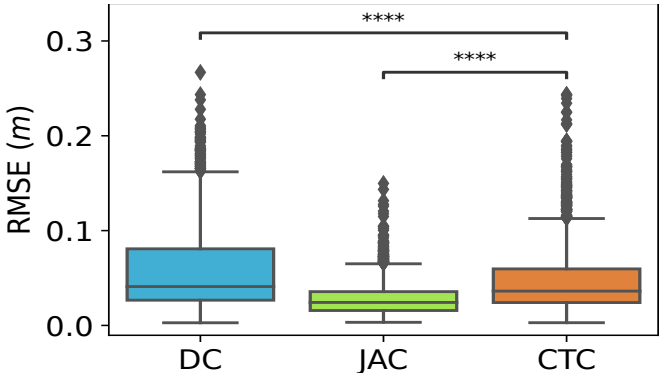


Fig. 2. Box plot for the RMSE on cursor position of all movements for the three considered cost functions.

expected. However, by far the best fit is obtained by joint acceleration costs (6), which we focus on in the following.

4.2 Effect of the MPC Horizon

Black box models like the one in our case suffer from the lack of efficiently available gradient information, resulting in considerably long optimization times. One way to handle extensive computation time is to reduce the MPC horizon. While this effectively reduces the number of control variables in each subproblem, the resulting MPC solution may further deviate from the (theoretic) optimal closed-loop trajectory for the infinite time-horizon problem (Grüne and Pannek, 2017). Therefore, the MPC horizon should be chosen carefully. To determine a suitable horizon length, we investigate our model with costs (6) and optimal cost weights r_1, r_2 from Section 4.1 by running a qualitative comparison for MPC horizons $N = 2, 4, 6, ..., 20$ on a representative trial of a single user.

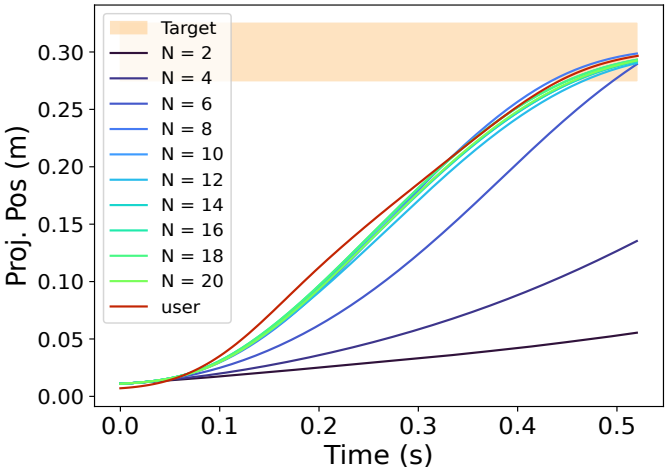


Fig. 3. Projected cursor position profiles of one trial for varying MPC horizon N, using JAC.

The resulting projected cursor trajectories[6] for the different MPC horizons are shown in Figure 3. As expected, a too short MPC horizon of $N = 2$ (80 ms) or $N = 4$ (160 ms) is not sufficient to reach the target. For $N = 6$ (200 ms), the controller manages to identify a control sequence that reaches the target, however, the simulation

[6] We orthogonally project the 3D trajectories onto the direct path between initial position and target.

clearly deviates from the user behavior (red line). For horizons starting from $N = 8$ (320 ms), trajectories hardly differ and visually match the user trajectory quite well.

4.3 Cost Function Sensitivity Analysis

To get insight in the individual effect of each cost weight r_1, r_2 in (6), we performed a numerical sensitivity analysis. For several combinations of the control weight r_1 and the joint acceleration weight r_2, we simulated all trials of a single participant in a single condition (61 in total) [7] and calculated the average RMSE between the corresponding simulation and user cursor position trajectories, analogous to (8).

Figure 4 shows an almost convex surface, indicating a local minimum close to $(r_1, r_2) = (0.01, 0.0001)$. This clearly shows that both too high control costs and too low joint acceleration costs result in large RMSE values. However, it is also visible that removing control costs completely, i.e., setting $r_1 = 0$, leads to a (slight) increase in RMSE values. This reflects that moving the human body does require some muscle control effort.

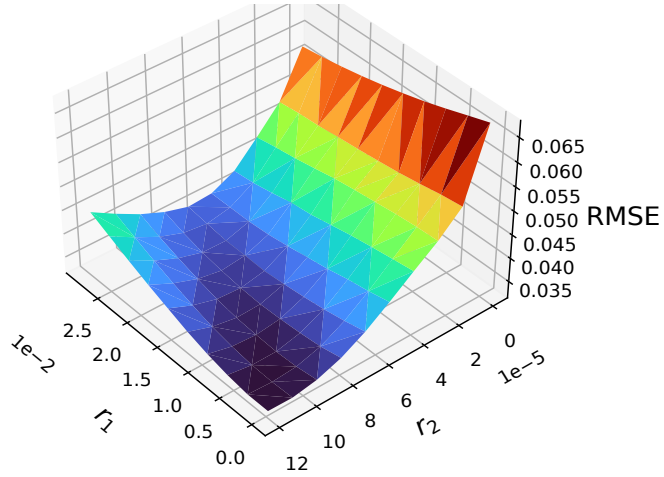


Fig. 4. Average RMSE for one representative participant and condition for different cost weights of JAC.

5. CONCLUSION

In this work, we demonstrated the application of Model Predictive Control to the HCI problem of evaluating interaction techniques through simulation. From the three considered cost functions, the combination of distance, control and joint acceleration cost showed the most promising results in terms of matching human pointing behavior. Our analysis of the MPC horizon showed that for the considered pointing task, a horizon of at least 320 ms (or $\approx 60\%$ of the total movement duration) is required to reliably reach the target. The subsequent sensitivity analysis brought novel insights in the choice of cost weights, showing a clear valley towards a (local) minimum, suggesting that large control costs should be avoided. With this work, we have established a sound basis for evaluating interaction techniques through simulation and have thus taken a first step towards being able to partially replace user studies in the future.

[7] We exemplarily chose participant U1 and the Virtual Cursor identity, i.e., the fingertip corresponds to the cursor position.

REFERENCES

Albrecht, S. (2013). *Modeling and numerical solution of inverse optimal control problems for the analysis of human motions*. Dissertation, Technische Universität München, München.

Berret, B., Chiovetto, E., Nori, F., and Pozzo, T. (2011). Evidence for composite cost functions in arm movement planning: An inverse optimal control approach. *PLOS Computational Biology*, 7(10), 1–18.

Diedrichsen, J., Shadmehr, R., and Ivry, R.B. (2010). The coordination of movement: optimal feedback control and beyond. *Trends in cognitive sciences*, 14(1), 31–39.

Fischer, F., Bachinski, M., Klar, M., Fleig, A., and Müller, J. (2021). Reinforcement learning control of a biomechanical model of the upper extremity. *Scientific Reports*, 11(1), 1–15.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6), 381.

Gawthrop, P., Loram, I., Lakie, M., and Gollee, H. (2011). Intermittent control: A computational theory of human control. *Biological Cybernetics*, 104(1-2), 31–51.

Grüne, L. and Pannek, J. (2017). *Nonlinear Model Predictive Control. Theory and Algorithms*. Springer, London, 2nd edition.

Hansen, N. (2016). The cma evolution strategy: A tutorial.

Müller, J. (2017). Dynamics of pointing with pointer acceleration. In *Human-Computer Interaction – INTERACT 2017*, 475–495. Springer International Publishing, Cham.

Müller, J., Oulasvirta, A., and Murray-Smith, R. (2017). Control theoretic models of pointing. *ACM Transactions on Computer-Human Interaction*, 24(4), 1–36.

Nakano, E., Imamizu, H., Osu, R., Uno, Y., Gomi, H., Yoshioka, T., and Kawato, M. (1999). Quantitative examinations of internal representations for arm trajectory planning: minimum commanded torque change model. *Journal of Neurophysiology*, 81(5), 2140–2155.

Saul, K.R., Hu, X., Goehler, C.M., Vidt, M.E., Daly, M., Velisar, A., and Murray, W.M. (2014). Benchmarking of dynamic simulation predictions in two software platforms using an upper limb musculoskeletal model. *Computer methods in biomechanics and biomedical engineering*, 5842(May 2016), 1–14.

Seth, A., Hicks, J.L., Uchida, T.K., Habib, A., Dembia, C.L., Dunne, J.J., Ong, C.F., DeMers, M.S., Rajagopal, A., Millard, M., et al. (2018). Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS computational biology*, 14(7), e1006223.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033.

van Beers, R.J., Haggard, P., and Wolpert, D.M. (2004). The role of execution noise in movement variability. *Journal of Neurophysiology*, 91(2), 1050–1063.

Van der Helm, F.C. and Rozendaal, L.A. (2000). Musculoskeletal systems with intrinsic and proprioceptive feedback. In *Biomechanics and neural control of posture and movement*, 164–174. Springer.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Machine Learning Enhanced Algorithm for Optimal Landing Problem

Yaohua Zang [*] **Jihao Long** [**] **Xuanxi Zhang** [***] **Wei Hu** [**]
Weinan E [***,**] **Jiequn Han** [****]

[*] *Zhejiang University, Hangzhou, Zhejiang 310027, China*
[**] *Princeton University, Princeton, NJ 08544, USA*
[***] *Peking University, Beijing 100871, China*
[****] *Flatiron Insitute, New York, NY 10010, USA (e-mail: jiequnhan@gmail.com)*

Abstract: We study the optimal landing problem for aerial vehicles under (1) a fixed landing time horizon or (2) the minimum time horizon. Both problems can be framed into solving the corresponding two-point boundary value problems. However, solving the boundary value problem in numerics is challenging, primarily due to the lack of good initial conditions. We present a space-marching scheme combined with machine learning techniques to provide good initial conditions for the boundary value problem solver. The algorithm greatly improves the solver's performance by increasing the success rate and reducing the computation time.

Keywords: optimal control, deep neural networks, Pontryagin minimum principle, quadrotors, landing problem

1. INTRODUCTION

The optimal landing problem concerns optimally controlling the aerial vehicles to land on the target position. Due to the high dimensionality of the state space and nonlinearity of the dynamics, solving the optimal landing problem has been a numerically challenging task for a long time. There has been a lot of work (Hu and Mishra, 2017; Zhu et al., 2019) focusing on solving the optimal landing problem. However, the exsiting methods still have some limitations such as simplified model, sub-optimal trajectory, reliance on good initial guesses, and long computation time.

In recent years, deep neural networks (DNN) have been widely used to solve challenging optimal control problems. For example, Zhu et al. (2019) use DNNs to learn the optimal action to improve the efficiency for the fuel-optimum lunar landing problem. Nakamura-Zimmerer et al. (2021) propose to learn the value function via DNN adaptively and predict the optimal feedback control.

This paper proposes a new numerical method enhanced by DNNs to solve the optimal landing problems and tackle the aforementioned difficulties. We will take the quadrotor unmanned aerial vehicles (UAVs) as an example to demonstrate our methodology. Quadrotor UAVs have received widespread attention in recent years due to their wide range of application scenarios. Our considered problem deals with the full quadrotor dynamic model and aims to achieve an optimal landing path with minimum time and control effort. We use the Pontryagin minimum principle to transform the original optimal landing problem into a two-point boundary value problem (TPBVP). One critical difficulty of TPBVP-based algorithms is to find good initial guesses. To overcome this difficulty, we design a DNN-based algorithm to provide an initial guess of the optimal landing time and a space-marching scheme to provide an initial guess of the control. Compared to the baseline methods, the proposed algorithm obtains the optimal landing trajectory with a much higher success rate and less computation time.

2. FORMULATION OF THE OPTIMAL CONTROL PROBLEM

We consider a deterministic system defined by the following ordinary differential equation

$$\begin{cases} \dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t), \boldsymbol{u}(t)), \quad t \in [0, t_f] \\ \boldsymbol{x}_0 = x_0, \ \boldsymbol{g}(\boldsymbol{x}(t_f)) = \boldsymbol{0}. \end{cases} \tag{1}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^n$ indicates the states, $\boldsymbol{u}(t) \in \mathcal{U} \subset \mathbb{R}^m$ represents the controls with $\mathcal{U}$ being the admissible set of the controls, $\boldsymbol{f} : \mathbb{R}^n \times \mathcal{U} \mapsto \mathbb{R}^n$ and $\boldsymbol{g} : \mathbb{R}^n \mapsto \mathbb{R}^k$ are smooth functions describing the dynamics and terminal condition. We call $\{\boldsymbol{x}, \boldsymbol{u}, t_f\}$ a feasible path if (1) is satisfied and use $\mathcal{P}$ to denote the set of all feasible paths. The *performance function* is defined by

$$J[\boldsymbol{x}, \boldsymbol{u}, t_f] = \int_0^{t_f} L(\boldsymbol{x}(t), \boldsymbol{u}(t))dt, \tag{2}$$

where $L : \mathbb{R}^n \times \mathcal{U} \mapsto \mathbb{R}^+$ is the nonnegative running cost.

We will consider two different but closely related problems. In the first problem, t_f is a given positive constant and we aim to minimize the performance function over all feasible paths with a fixed terminal time t_f:

$$\min_{(\boldsymbol{x}, \boldsymbol{u}) : \{\boldsymbol{x}, \boldsymbol{u}, t_f\} \in \mathcal{P}} J[\boldsymbol{x}, \boldsymbol{u}, t_f]. \tag{3}$$

We call this problem a *fixed terminal time problem*. Another problem is called *free terminal time problem*, where

we aim to minimize the performance function over all feasible paths:

$$\min_{\{\boldsymbol{x},\boldsymbol{u},t_f\}\in\mathcal{P}} J[\boldsymbol{x},\boldsymbol{u},t_f]. \tag{4}$$

2.1 Pontryagin's Minimum Principle

Pontryagin's minimum principle (PMP) derives the necessary conditions for optimality, which converts the optimal control problems (3) or (4) to two-point boundary value problems (TPBVPs). In this paper, we assume the solutions of the TPBVP are optimal. Then one can solve the TPBVPs to obtain the solution to the original control problem. To state the PMP, we introduce the costate variable $\lambda \in \mathbb{R}^n$ and define the Hamiltonian function

$$H(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{u}) = L(\boldsymbol{x},\boldsymbol{u}) + \boldsymbol{\lambda}\cdot f(\boldsymbol{x},\boldsymbol{u}). \tag{5}$$

The PMP reduces the fixed terminal time problem (3) to a system of $2n$ equations in the form of a TPBVP

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \partial^T_{\boldsymbol{\lambda}} H(\boldsymbol{x}(t),\boldsymbol{\lambda}(t),\boldsymbol{u}^*(t)), \\ \dot{\boldsymbol{\lambda}}(t) = \partial^T_{\boldsymbol{x}} H(\boldsymbol{x}(t),\boldsymbol{\lambda}(t),\boldsymbol{u}^*(t)), \end{cases} \tag{6}$$

with the boundary conditions given by the original boundary conditions augmented with the transversality conditions:

$$\begin{cases} \boldsymbol{g}(\boldsymbol{x}(t_f)) = \boldsymbol{0}, \\ \mathcal{A}\nabla\boldsymbol{g}(\boldsymbol{x}(t_f)) = \boldsymbol{\lambda}(t_f). \end{cases} \tag{7}$$

Here $\mathcal{A} \in \mathbb{R}^{n\times k}$ is a matrix multiplier. The optimal control $\boldsymbol{u}^*(t)$ should satisfy the minimization of the Hamiltonian at each t:

$$\boldsymbol{u}^*(t) = \arg\min_{\boldsymbol{u}\in\mathcal{U}} H(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{u}). \tag{8}$$

Equations (6), (7) and (8) together complete the PMP for the fixed terminal time problem (3). For the free terminal time problem (4), besides (6), (7) and (8), we need the following condition to determinate the optimal terminal time t_f:

$$H(\boldsymbol{x}(t_f),\boldsymbol{\lambda}(t_f),\boldsymbol{u}^*(t_f)) = 0. \tag{9}$$

2.2 Optiaml landing problem

The dynamics of the quadrotor UAV is modeled as follows

$$\begin{cases} \dot{\boldsymbol{p}} = \boldsymbol{R}^T(\boldsymbol{\eta})\boldsymbol{v}_b \\ \dot{\boldsymbol{v}}_b = -\boldsymbol{w}_b\times\boldsymbol{v}_b - \boldsymbol{R}(\boldsymbol{\eta})\boldsymbol{g} + \dfrac{1}{m}\boldsymbol{f}_u \\ \dot{\boldsymbol{\eta}} = \boldsymbol{K}(\boldsymbol{\eta})\boldsymbol{w}_b \\ \dot{\boldsymbol{w}}_b = -\boldsymbol{J}^{-1}\boldsymbol{w}_b\times\boldsymbol{J}\boldsymbol{w}_b + \boldsymbol{J}^{-1}\boldsymbol{\tau}_u, \end{cases} \tag{10}$$

where $\boldsymbol{p} = (x,y,z)^T$ is the inertial position in the earth frame and $\boldsymbol{v}_b = (v_x,v_y,v_z)$ is the inertial velocity of the quadrotor in the body frame. $\boldsymbol{\eta} = (\phi,\theta,\psi)$ is the attitude of the quadrotor in the earth frame defined by the Euler angles: roll(ϕ), pitch(θ) and yaw(ψ). $\boldsymbol{w}_b = (p,q,r)^T$ denotes the angular velocity in the body frame. $\boldsymbol{f}_u = (0,0,T)^T$ and $\boldsymbol{\tau}_u = (\tau_x,\tau_y,\tau_z)^T$ are the total thrust and body torques from four rotors, which are forces applied by the control variables to adjust the quadrotor's dynamics. The constants m and $\boldsymbol{g} = (0,0,g)^T$ denote the mass and the gravity vector, respectively. $\boldsymbol{R}(\boldsymbol{\eta}) \in SO(3)$ and $\boldsymbol{K}(\boldsymbol{\eta})$ are given matrix functions denoting the direction cosine matrix and attitude kinematic matrix. $\boldsymbol{J}$ denotes the constant inertia matrix. We let $\boldsymbol{x} = (\boldsymbol{p}^T,\boldsymbol{v}_b^T,\boldsymbol{\eta}^T,\boldsymbol{w}_b^T)^T \in \mathbb{R}^{12}$ so that $\boldsymbol{x}$ denotes the state variable of our optimal landing problem. Meanwhile, we denote the control as $\boldsymbol{u} = (T,\tau_x,\tau_y,\tau_z)^T$. Then we have $\boldsymbol{f}_u = A\boldsymbol{u}$ and $\boldsymbol{\tau}_u = B\boldsymbol{u}$ with A and B are two constant matrices.

We aim to solve the quadrotor landing problem with minimum time and control effort under the dynamics described above. That is, to find the optimal controls to steer the UAVs from some initial states $\boldsymbol{x}_0 \in \mathcal{S}_0$ to a given target $\boldsymbol{x}_{t_f} \in \mathcal{S}_{t_f}$ satisfying $\boldsymbol{g}(\boldsymbol{x}_{t_f}) = \boldsymbol{0}$. For the landing problem, $\mathcal{S}_{t_f}$ has the form $\{\boldsymbol{x}_{t_f} \mid \boldsymbol{p}(t_f) = \boldsymbol{v}(t_f) = \boldsymbol{w}(t_f) = \boldsymbol{0}; \phi(t_f) = \theta(t_f) = 0\}$. The running cost L is defined as

$$L(\boldsymbol{x},\boldsymbol{u}) = 1 + (\boldsymbol{u}-\boldsymbol{u}_d)^T Q_u(\boldsymbol{u}-\boldsymbol{u}_d),$$

where $\boldsymbol{u}_d = (mg,0,0,0)$ represents the reference control that balances with gravity and $Q_u = \mathrm{diag}(1,1,1,1)$ represents the weight matrix characterizing the cost of deviating from the reference control. Then the description of the optimal landing problem is complete. With the aforementioned Pontryagin's minimum principle, we can transform both optimal landing problems with fixed terminal time and free terminal time to the corresponding TPBVPs.

3. MACHINE LEARNING ENHANCED METHOD FOR SOLVING OPTIMAL LANDING PROBLEM

This section presents our machine learning enhanced algorithm for solving the optimal landing problem, accompanied by the numerical results on the optimal landing problem of the quadrotor UAVs. The same system parameters as in Madani and Benallegue (2006) are used in this paper: the mass $m = 2kg$, the gravity $g = 9.81m/s^2$, the moment of of inertia $J_x = J_y = J_z/2 = 1.2416kg\cdot m^2$. We specify the domain of initial state as $\mathcal{S}_0 = \{x,y \in [-10,10], z \in [5,100], v_x,v_y,v_z \in [-0.5,0.5], \theta,\phi \in [-\pi/4,\pi/4], \psi \in [-\pi,\pi]; \boldsymbol{w} = \boldsymbol{0}\}$. We always uniformly sample 100 initial positions $\boldsymbol{x}_0$ from $\mathcal{S}_0$ to estimate the success rate and average computation time of the algorithm. Throughout the paper, we will use the BVP solver in Kierzenka and Shampine (2001) to solve the TPBVP. Figure 1 presents an example of the optimal landing path of the quadrotor UAV from the starting position $\boldsymbol{x}_0$ to the origin.

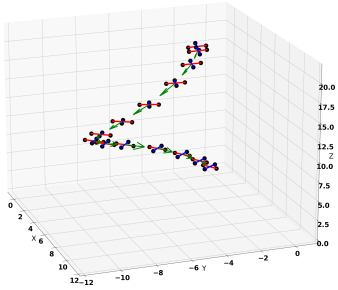


Fig. 1. An example of the optimal trajectory for the landing problem

3.1 Use Solution to the fixed terminal time as initial guess

Although we have the PMP (6)–(9) characterizing the solution of the optimal landing problem with free terminal time, we observe that the associated TPBVP is numerically challenging to solve. The main reason is the lack of a good initial guess for t_f and the paths of $\boldsymbol{x}(t),\boldsymbol{\lambda}(t)$. Without any prior knowledge, the simplest choice is to

initialize t_f with a reasonable scalar and $\boldsymbol{x}(t), \boldsymbol{\lambda}(t)$ with constant zero. However, the BVP solver hardly converges under this choice. Table 1 reports its success rate with a few different initial guesses of t_f. We can see that, with zero initialization of the path, the success rate of solving free terminal time TPBVP is always extremely low, regardless of the initial guess of t_f.

Table 1. Solving TPBVP corresponding to the free terminal time problems with zero initialization

initial guess $\tilde{t}_f^*$	4	8	12	16	20	24
success rate	3%	4%	0%	0%	1%	1%

Table 1 suggests that the initial guess of $\boldsymbol{x}(t)$ and $\boldsymbol{\lambda}(t)$ plays an important role in the convergence of the BVP solver when solving the free terminal time problem. To address this issue, we notice that the solution of the free terminal time problem is also the solution of a corresponding fixed terminal time problem if the fixed terminal time t_f equals the optimal terminal time t_f^*. In other words, if we have a reasonable guess of t_f^*, the solution of the fixed terminal time problem can provide us a good initial guess to the free terminal time problem. Moreover, the fixed terminal time problem is easier to solve with many efficient techniques, such as marching methods introduced in the next subsection. Therefore, we can first guess a value of the optimal terminal time $\tilde{t}_f^*$ and solve the fixed terminal time problem with $t_f = \tilde{t}_f^*$. Then we use its solution as the initial guess to solve the free terminal time problem. This approach can be viewed as a warm start method for solving the optimization problem. The corresponding algorithm is summarized in Algorithm 1, in which we again use zero to initialize the paths. The corresponding numerical results are presented in Figure 2. In Figure 2, the label "free" means the success rate of the free terminal time problem while the label "fix/free" means the success rate of problems that the fix terminal time problem has been solved while the corresponding free terminal time problem failed. Because we will not solve the free terminal time problem when the fixed terminal time problem fails, the success rate of the free terminal time problem is always lower than the success rate of the fixed terminal time problem. Comparing Figure 2 with Table 1, we can see that warming start with the solution to the fixed terminal time problem significantly improves the success rate, although the rate is still not high enough for practical applications.

Algorithm 1 Warm start with fixed terminal time solution

1: **Input:** The initial state x_0; the guess value of the optimal terminal time $\tilde{t}_f^*$.
2: Solve the fixed time problem with $t_f = \tilde{t}_f^*$ using zero initialization.
3: Solve the free time problem by using the solution of the corresponding fixed time problem as initial guess.
4: **Output:** The solution of the free time problem.

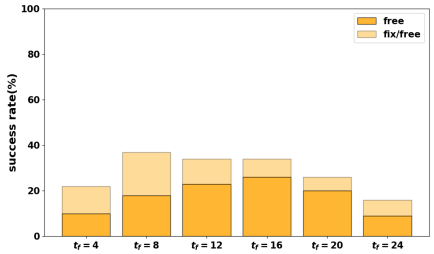


Fig. 2. Sucess rate of Algoriuthm 1: warm start with fixed terminal time solution.

3.2 Space-marching for solving the fixed terminal time problem

To further improve the success rate of solving the TPBVP corresponding to the fixed terminal time problem, we propose a space-marching method (Murio, 2002) tailored to the optimal landing problem. Its intuition is as follows. Solving the fixed time problem is still difficult since the initial state $\boldsymbol{x}_0$ is far away from the terminal set $\mathcal{S}_{t_f}$. We can solve a simpler fixed time problem in which the initial state is closer to the terminal state, while the solution is not far from the original fixed time problem. After the simpler fixed time problem is solved, we can use its solution as the initial guess to solve the original harder one. To present this method in a more general and systematic way, we say $\boldsymbol{x}_{end}$ is a terminal state if there exists $\boldsymbol{u} \in \mathcal{U}$ such that for any $t_f \geq 0$, the path

$$\boldsymbol{x}(t) \equiv \boldsymbol{x}_{end},\ \boldsymbol{u}(t) \equiv \boldsymbol{u},\ 0 \leq t \leq t_f$$

is the optimal path for the fixed-time problem with t_f as terminal time and $\boldsymbol{x}_0 = \boldsymbol{x}_{end}$. We always assume such a terminal state exists for the optimal landing problem. In this paper, we choose the origin point (of the 12-dimensional state space) as the terminal state. Then we evenly select K points in the line segment from $\boldsymbol{x}_{end}$ to $\boldsymbol{x}_0$, and denote them as $\{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \cdots, \boldsymbol{x}_0^K\}$ according to their increasing distances to $\boldsymbol{x}_{end}$ ($\boldsymbol{x}_0^K = \boldsymbol{x}_0$). In each marching step, we solve the fixed time problem with the initial state $\boldsymbol{x}_0^k$ by using the solution obtained from the fixed time problem with the initial state $\boldsymbol{x}_0^{k-1}$ as the initial guess. The process repeats until $k = K$. This algorithm is called warm start with fixed terminal time solution through space-marching, as summarized in Algorithm 2.

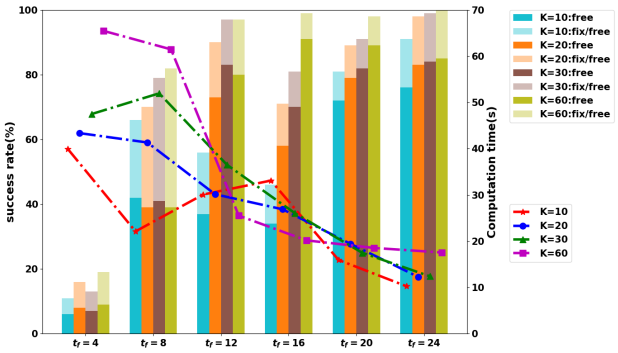


Fig. 3. Success rate and computation time of Algorithm 2: warm start with fix terminal time solution through space-marching with different K.

Algorithm 2 Warm start with fixed terminal time solution through space-marching

1: **Input:** The initial state $\boldsymbol{x}_0$; the guess value of the optimal terminal time $\tilde{t}_f^*$; the marching number K.
2: Evenly select K points in the line segment from $\boldsymbol{x}_{end}$ to $\boldsymbol{x}_0$, and denotes them as $\{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \cdots, \boldsymbol{x}_0^K\}$.
3: Set the initial guess $\boldsymbol{x}_{aug}$ as the zero initial guess.
4: **for** $k = 1, 2, \cdots, K$ **do**
5: Solve the fixed terminal time problem with initial state $\boldsymbol{x}_0^k$ by using the initial guess $\boldsymbol{x}_{aug}$.
6: replace the initial guess $\boldsymbol{x}_{aug}$ with the solution.
7: **end for**
Solving the free time problem with $\boldsymbol{x}_{aug}$ as the initial guess of the path and $\tilde{t}_f^*$ as the initial guess of terminal time.
8: **Output:** The solution of the free terminal time problem with initial state $\boldsymbol{x}_0$.

Figure 3 shows the success rate and computation time of Algorithm 2 with different choices of the initial guess $\tilde{t}_f^*$ and K. We can see that most fixed terminal time problems can be solved with a high success rate if the predicted terminal time $\tilde{t}_f^*$ is not too small and the marching number K is large enough. However, the free terminal time problem may not be solved successfully if the guessed terminal time is not close enough to the real terminal time. To further improve the success rate of the free terminal time problem, we need a more accurate prediction of the optimal terminal time.

3.3 Predict the Optimal Terminal Time

Now we consider empowering Algorithm 2 by predicting the optimal terminal time as a function of the initial state $\boldsymbol{x}_0$ through a linear model or a neural network. To do so, we first prepare a dataset for supervised learning. We randomly select 300 initial positions $\boldsymbol{x}_0$ and use Algorithm 3 with $K = 60, \tilde{t}_f^* = 24$ to collect 300 optimal landing paths. We select 100 positions (uniformly in time) on each optimal landing path and store the corresponding optimal landing time to obtain the training data. We have 30000 pairs of starting positions and optimal ending times for training in total. We then use them to optimize a linear model and a neural network model (3 three hidden layers and 64 neurons in each layer) based on the objective being the squared difference between the predicted $\tilde{t}_f^*$ and the truth optimal terminal time. Then, when we need to solve a free terminal time problem with a new initial state $\boldsymbol{x}_0$, we first use the linear model or the neural network to predict the optimal terminal time $\tilde{t}_f^*$ associated with $\boldsymbol{x}_0$ and use Algorithm 2 to solve the free terminal time problem.

The success rates of using a linear model or a neural network to predict the optimal terminal time with different space-marching numbers are presented in Figure 4. Comparing the results with those using the guessed constant optimal terminal time independent of the initial state in Figure 3, we can see that both the linear model and neural network model achieves much higher success rates. Using a neural network attains higher success rates and takes less computation time because it can more accurately predict the optimal terminal time. With the help of neural networks and the space-marching with $K = 60$, we achieve a 99% successful rate and the average computational time is about 17 seconds, which performs the best among the methods considered in this paper.

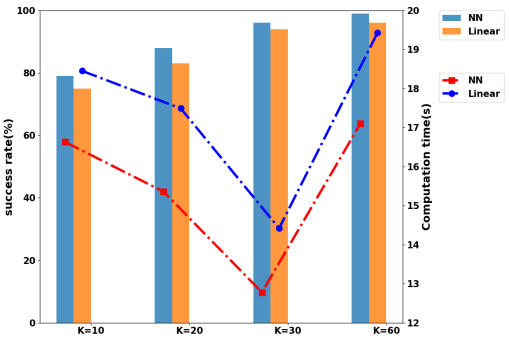


Fig. 4. NN prediction of t_f as the input to Algorithm 2 vs. Linear prediction of t_f as the input to Algorithm 2.

4. SUMMARY AND DISCUSSION

In this paper, we propose a machine learning enhanced method to solve the optimal control problem of quadrotor landing. To solve the free terminal TPBVP, we first solve a fixed terminal time TPBVP and then use its solution as an initial guess. The initial guess for the optimal landing time is predicted by a neural network that is trained in advance. A space-marching method is proposed to improve the efficiency of solving the fixed terminal time problem. We demonstrate the effectiveness of the proposed method through a series of experiments.

When solving the optimal landing problem, we have observed that directly learning the mapping from states to controls or value functions with a DNN and using its prediction as an initial guess for the TPBVP leads to divergence. Improving the stability of neural network-based control is an important future direction of research.

REFERENCES

Hu, B. and Mishra, S. (2017). A time-optimal trajectory generation algorithm for quadrotor landing onto a moving platform. In *2017 American Control Conference (ACC)*, 4183–4188. IEEE.

Kierzenka, J. and Shampine, L.F. (2001). A BVP solver based on residual control and the Maltab PSE. *ACM Transactions on Mathematical Software (TOMS)*, 27(3), 299–316.

Madani, T. and Benallegue, A. (2006). Backstepping control for a quadrotor helicopter. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3255–3260. IEEE.

Murio, D.A. (2002). Mollification and space marching. *Inverse engineering handbook*, 219–326.

Nakamura-Zimmerer, T., Gong, Q., and Kang, W. (2021). Adaptive deep learning for high-dimensional Hamilton–Jacobi–Bellman equations. *SIAM Journal on Scientific Computing*, 43(2), A1221–A1247.

Zhu, L., Ma, J., and Wang, S. (2019). Deep neural networks based real-time optimal control for lunar landing. In *IOP Conference Series: Materials Science and Engineering*, volume 608, 012045. IOP Publishing.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Orchestrating front and rear sensors for global stabilization of unicycles

Riccardo Ballaben * **Philipp Braun** ** **Luca Zaccarian** ***

* *University of Trento (e-mail: riccardo.ballaben@studenti.unitn.it)*
** *Australian National University (e-mail: philipp.braun@anu.edu.au)*
*** *LAAS-CNRS Université de Toulouse, CNRS, Toulouse, France and University of Trento, Italy (e-mail: luca.zaccarian@laas.fr)*

Abstract: We consider mobile robots described through unicycle dynamics equipped with range sensors and cameras, one in the front and one in the back providing measurements of the distance and misalignment to a target. We derive locally asymptotically stabilizing control laws driving the robot to the target position and orientation. The local control laws are combined into a hybrid global stabilizer, switching between control laws relying on the measurements from the front and rear sensors. Using Lyapunov arguments in the local setting as well as in the hybrid systems formulation, we prove global asymptotic stability of the target set for the hybrid closed-loop system. The results are illustrated on numerical examples.

1. INTRODUCTION & MOTIVATION

Driving a robot described through unicycle dynamics to a target set with a particular fixed final orientation is a difficult task due to nonholonomic constraints. In particular, the origin of the unicycle dynamics cannot be globally asymptotically stabilized through a static state feedback Brockett (1983). Indeed unicycle dynamics do not satisfy the so-called Brockett conditions. Control laws guaranteeing convergence to the origin, thus imply the necessity to combine the controller designs with reference tracking or path following approaches or to rely on discontinuous feedback laws instead. We refer to Tzafestas (2013) as a general reference for mobile robots and control.

In this work, we follow the second path, i.e., we consider discontinuous feedback laws. While Lipschitz-continuous feedback laws guarantee some intrinsic robustness properties with respect to stability and with respect to existence and uniqueness of solutions, well-posedness and robustness is more difficult to achieve with discontinuous feedback laws (Sontag (1999)). To define a globally asymptotically stabilizing feedback we use a hybrid systems formalism and borrow results from hybrid Lyapunov theory (Goebel et al. (2012)). We define locally stabilizing control laws and orchestrate them through a switching mechanism to obtain global results. The local controller design is motivated through the results and derivations in Aicardi et al. (1995), whereas the global design and the setup are motivated and derived differently.

We consider mobile robots, which in Cartesian coordinates are described through the dynamics

$$\dot{x} = \begin{bmatrix} \dot{p}_1 \\ \dot{p}_2 \\ \dot{\phi} \end{bmatrix} = f(x,u) = \begin{bmatrix} u_1 \cos(\phi) \\ u_1 \sin(\phi) \\ u_2 \end{bmatrix}, \tag{1}$$

where $p = [p_1 \ p_2]^T \in \mathbb{R}^2$ captures the unicycle position in the plane, $\phi \in \mathbb{R}$ captures its orientation and the input $u = [u_1 \ u_2]^T \in \mathbb{R}^2$ captures the velocity and angular velocity.

⋆ P. Braun and L. Zaccarian are supported in part by the Agence Nationale de la Recherche (ANR) via grant "Hybrid And Networked Dynamical sYstems" (HANDY), number ANR-18-CE40-0010.

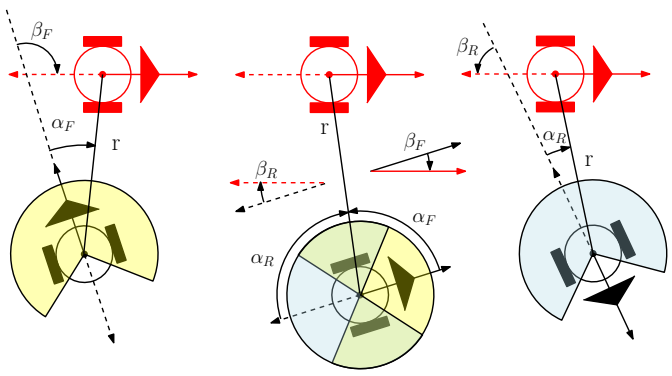


Fig. 1. The robot is equipped with range sensors and cameras (a front and a rear camera) with overlapping fields of view, providing measurements $(r, \alpha_\mathrm{F}, \beta_\mathrm{F})$ and $(r, \alpha_\mathrm{R}, \beta_\mathrm{R})$, respectively. Each measurement is only available when the target is in the camera's field of view $\alpha_i \in [-\frac{\pi}{2} - \delta, \frac{\pi}{2} + \delta]$, $i \in \{\mathrm{F}, \mathrm{R}\}$.

We assume that the mobile robot is equipped with various sensors, including a range sensor, measuring the distance $r \in \mathbb{R}_{\geq 0}$ to the target position, and two cameras (a front and a rear camera). The cameras provide measurements of the misalignment of the heading of the robot in terms of angles α_F and α_R corresponding to measurements from the front and the rear camera. In addition, the cameras provide measurements of the angles β_F and β_R, as represented in Figure 1. By combining α_i and β_i, $i \in \{\mathrm{F}, \mathrm{R}\}$, the mismatch of the robot orientation and the target orientation is defined. The setting is visualized in Figure 1, where it is apparent that, the field of view of both front and rear cameras are as follows

$$\alpha_\mathrm{F}, \alpha_\mathrm{R} \in [-\tfrac{\pi}{2} - \delta, \tfrac{\pi}{2} + \delta], \tag{2}$$

where $\delta \in (0, \frac{\pi}{2})$ induces some overlap and ensures that the combined fields of view are covering an area of 360°. While in certain configurations only one camera is available (left and right cases in Figure 1), in some configurations the target is in the field of view of both cameras. This motivates the use of a hybrid controller, which switches between measurements from the two cameras and makes

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

use of different error dynamics describing the mismatch of the robot and the target position and orientation.

2. ROBOT DYNAMICS

In this section we derive dynamics in local coordinates defined as $z_i = [r_i \ \beta_i \ \alpha_i]^T$, $i \in \{\mathrm{R}, \mathrm{F}\}$. The relation between the mobile robot in Cartesian coordinates and the local coordinates related to the sensor measurements, is described through the coordinate transformations

$$\begin{bmatrix} p_1 \\ p_2 \\ \phi \end{bmatrix} = \begin{bmatrix} r\cos(\alpha_\mathrm{R} - \beta_\mathrm{R}) \\ r\sin(\alpha_\mathrm{R} - \beta_\mathrm{R}) \\ -\beta_\mathrm{R} \end{bmatrix}, \quad \begin{bmatrix} p_1 \\ p_2 \\ \phi \end{bmatrix} = \begin{bmatrix} r\cos(\alpha_\mathrm{F} - \beta_\mathrm{F} - \pi) \\ r\sin(\alpha_\mathrm{F} - \beta_\mathrm{F} - \pi) \\ -\beta_\mathrm{F} \end{bmatrix} \tag{3}$$

where F and R again correspond to the front and rear cameras. The coordinate transformations (3) follow from trigonometric arguments applied to the variables defined and illustrated through Figure 2.

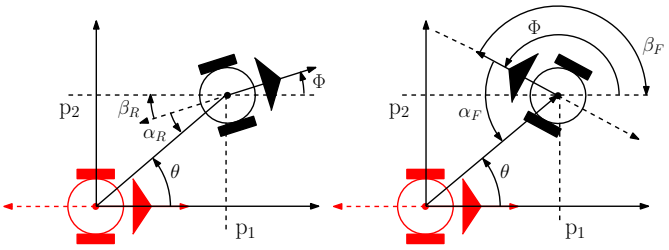


Fig. 2. Coordinate transformation from global Cartesian coordinates to local coordinates depending on the robot orientation.

The local coordinates are defined on the domain

$$z_i \in \mathcal{Z} := \mathbb{R}_{\geq 0} \times \mathbb{R} \times [-\tfrac{\pi}{2} - \delta, \tfrac{\pi}{2} + \delta], \quad i \in \{\mathrm{F}, \mathrm{R}\}, \tag{4}$$

for $\delta \in (0, \frac{\pi}{2})$, where the third component of z_i evolves in a bounded set as defined in (2). Whenever the target is in the field of view of both cameras (see the middle sketch in Figure 1), by simple geometric considerations we obtain

$$\begin{aligned} \alpha_\mathrm{F} &= \alpha_\mathrm{R} - \pi\,\mathrm{sign}(\alpha_\mathrm{R}), \qquad \beta_\mathrm{F} = \beta_\mathrm{R}, \\ \alpha_\mathrm{R} &= \alpha_\mathrm{F} - \pi\,\mathrm{sign}(\alpha_\mathrm{F}), \qquad \beta_\mathrm{R} = \beta_\mathrm{F}. \end{aligned} \tag{5}$$

Moreover, again due to the fact that $\delta > 0$ by assumption, for $r \neq 0$, i.e., $|p| \neq 0$, we can equivalently represent the robot through Cartesian coordinates x or through at least one of the local coordinates z_i, $i \in \{\mathrm{F}, \mathrm{R}\}$.

As a next step, we derive from (1) and (3) the dynamics in the local coordinates

$$\dot{z}_\mathrm{R} = f_\mathrm{R}(z_\mathrm{R}, v) \qquad \text{and} \qquad \dot{z}_\mathrm{F} = f_\mathrm{F}(z_\mathrm{F}, v), \tag{6}$$

with a transformed input $v \in \mathbb{R}^2$.

Lemma 1. Whenever $r \neq 0$, the dynamics (1) can be equivalently represented through (6) where

$$f_\mathrm{R}(z_\mathrm{R}, v) = \begin{bmatrix} v_1 r\cos(\alpha_\mathrm{R}) \\ -v_2 \\ -v_1\sin(\alpha_\mathrm{R}) - v_2 \end{bmatrix}, f_\mathrm{F}(z_\mathrm{F}, v) = \begin{bmatrix} -v_1 r\cos(\alpha_\mathrm{F}) \\ -v_2 \\ v_1\sin(\alpha_\mathrm{F}) - v_2 \end{bmatrix} \tag{7}$$

with transformed input $v_1 r = u_1$ and $v_2 = u_2$. ⌟

Proof. We start with a derivation of f_R. Let $r \neq 0$ for the remainder of the proof and define the angle $\theta = \alpha_\mathrm{R} - \beta_\mathrm{R}$.

Proceeding as in (Nešić et al., 2011, eq. (41)), we provide an alternative expression of the Jacobian $\frac{\partial p}{\partial p} = I$. In particular, using the definition of θ as well as (3), the matrix can be rewritten as

$$\begin{aligned} \frac{\partial p}{\partial p} &= \begin{bmatrix} \frac{\partial r}{\partial p_1}\cos(\theta) & \frac{\partial r}{\partial p_2}\cos(\theta) \\ \frac{\partial r}{\partial p_1}\sin(\theta) & \frac{\partial r}{\partial p_2}\sin(\theta) \end{bmatrix} + r \begin{bmatrix} -\sin(\theta)\frac{\partial \theta}{\partial p_1} & -\sin(\theta)\frac{\partial \theta}{\partial p_2} \\ \cos(\theta)\frac{\partial \theta}{\partial p_1} & \cos(\theta)\frac{\partial \theta}{\partial p_2} \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \begin{bmatrix} \frac{\partial r}{\partial p_1} & \frac{\partial r}{\partial p_2} \end{bmatrix} + r \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} \begin{bmatrix} \frac{\partial \theta}{\partial p_1} & \frac{\partial \theta}{\partial p_2} \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \frac{\partial r}{\partial p} + r \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} \frac{\partial \theta}{\partial p}. \end{aligned}$$

With the definition $J = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ and the observations that

$$\frac{p}{r} = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \quad \text{and} \quad Jp = r \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix}, \tag{8}$$

for $r \neq 0$, the calculations above can be further simplified and summarized as

$$I = \frac{\partial p}{\partial p} = \frac{\partial}{\partial p}\left(r \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}\right) = \frac{1}{r} p \frac{\partial r}{\partial p} + Jp \frac{\partial \theta}{\partial p}. \tag{9}$$

Note that $p^T J p = 0$ follows from the definition of J. Hence, left-multiplying (9) by p^T gives

$$p^T = \tfrac{1}{r} p^T p \tfrac{\partial r}{\partial p} + p^T J p \tfrac{\partial \theta}{\partial p} = \tfrac{r^2}{r} \tfrac{\partial r}{\partial p} = r \tfrac{\partial r}{\partial p}.$$

Similarly, the following equations hold

$$(Jp)^T I = \tfrac{1}{r} p^T J^T p \tfrac{\partial r}{\partial p} + p^T p \tfrac{\partial \theta}{\partial p} = r^2 \tfrac{\partial \theta}{\partial p},$$

which leads to the partial derivatives

$$\frac{\partial r}{\partial p} = \tfrac{1}{r} p^T \quad \text{and} \quad \frac{\partial \theta}{\partial p} = \tfrac{1}{r^2} (Jp)^T. \tag{10}$$

With these calculations, the definition of the function f_B follows from (1):

$$\begin{aligned} \dot{r} &= \tfrac{\partial r}{\partial p}\dot{p} = \tfrac{1}{r} p^T \dot{p} = \tfrac{1}{r} \begin{bmatrix} r\cos(\theta) & r\sin(\theta) \end{bmatrix} \begin{bmatrix} u_1\cos(\phi) \\ u_2\sin(\phi) \end{bmatrix} \\ &= u_1[\cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi)] = v_1 r\cos(\phi - \theta) \\ &= v_1 r\cos(-\beta_\mathrm{R} - \alpha_\mathrm{R} + \beta_\mathrm{R}) = v_1 r\cos(\alpha_\mathrm{R}) \\ \dot{\theta} &= \tfrac{\partial \theta}{\partial p}\dot{p} = \tfrac{1}{r^2} p^T J^T \dot{p} = \tfrac{1}{r^2} \begin{bmatrix} r\cos(\theta) & r\sin(\theta) \end{bmatrix} \begin{bmatrix} u_1\sin(\phi) \\ -u_1\cos(\phi) \end{bmatrix} \\ &= u_1(\cos(\theta)\sin(\phi) - \sin(\theta)\cos(\phi)) = v_1\sin(\phi - \theta)\alpha_\mathrm{R}) \\ &= -v_1\sin(\alpha_\mathrm{R}) \\ \dot{\beta}_\mathrm{R} &= -\dot{\phi} = -v_2 \\ \dot{\alpha}_\mathrm{R} &= \dot{\beta}_\mathrm{R} + \dot{\theta} = -v_2 - v_1\sin(\alpha_\mathrm{R}). \end{aligned}$$

Observe that for f_F we need to consider the relation $\alpha_\mathrm{R} = \alpha_\mathrm{F} - \pi$ and thus the representation f_F follows. □

While Lemma 1 excludes the case $r = 0$, note that the functions f_R and f_F are well-defined for $r = 0$.

3. LOCAL CONTROLLER DESIGN

In this section, we derive control laws locally asymptotically stabilizing the set

$$\mathcal{A} = \{0\} \times \mathbb{R} \times \{0\} \tag{11}$$

as well as the origin of the dynamical system (6). The Lyapunov construction is inspired by Aicardi et al. (1995).

Lemma 2. Let the feedback gains $k_r, k_\alpha \in \mathbb{R}_{>0}$ be arbitrary. Consider the dynamics (6) with $z_\mathrm{R} \in \mathcal{Z}$. Then

$$v_\mathrm{R} = \begin{bmatrix} -k_r\cos(\alpha_\mathrm{R}) \\ k_r\cos(\alpha_\mathrm{R})\sin(\alpha_\mathrm{R}) + k_\alpha\alpha_\mathrm{R} \end{bmatrix} \tag{12}$$

locally asymptotically stabilizes the set $\mathcal{A}$ in (11). Moreover, $V(z_\mathrm{R}) = \frac{1}{2}(r^2 + \alpha_\mathrm{R}^2)$ is a Lyapunov function for the closed-loop dynamics with respect to $\mathcal{A}$. ⌟

Proof. First observe that $\frac{1}{2}|z_\mathrm{R}|_\mathcal{A}^2 \leq V(z_\mathrm{R}) \leq \frac{1}{2}|z_\mathrm{R}|_\mathcal{A}^2$, i.e., V radially unbounded. Moreover, the directional derivative of V along the dynamics satisfies

$$\begin{aligned}\langle \nabla V(z_{\mathrm{R}}), & f_{\mathrm{R}}(z_{\mathrm{R}}, v_{\mathrm{R}})\rangle \\ &= r^2 v_{\mathrm{R}_1}\cos(\alpha_{\mathrm{R}}) + \alpha(-v_{\mathrm{R}_2} - v_{\mathrm{R}_1}\sin(\alpha_{\mathrm{R}})) \\ &= -k_r r^2\cos^2(\alpha_{\mathrm{R}}) - k_\alpha \alpha_{\mathrm{R}}^2 < 0,\end{aligned}$$

and thus $\frac{d}{dt}V(z_{\mathrm{R}}(t)) < 0$ for all $z_{\mathrm{R}} \in \mathbb{R}^3 \backslash \mathcal{A}$ which implies asymptotic stability of $\mathcal{A}$. □

To additionally ensure that $\beta_{\mathrm{R}}(t) \to 0$ for $t \to \infty$, we may include an additional term in the feedback law v_{R} and in the Lyapunov function.

Lemma 3. Let $k_r, k_\alpha, k_\beta \in \mathbb{R}_{>0}$ be arbitrary. Consider the dynamics (6) with $z_{\mathrm{R}} \in \mathcal{Z}$. Then the control law

$$v_{\mathrm{R}} = \begin{bmatrix} -k_r\cos(\alpha_{\mathrm{R}}) \\ k_r\cos(\alpha_{\mathrm{R}})\sin(\alpha_{\mathrm{R}}) + k_\alpha\alpha_{\mathrm{R}} + k_\beta(\alpha_{\mathrm{R}} - \beta_{\mathrm{R}})\frac{\cos(\alpha_{\mathrm{R}})\sin(\alpha_{\mathrm{R}})}{\alpha_{\mathrm{R}}} \end{bmatrix} \tag{13}$$

is well-defined for all $z_{\mathrm{R}} \in \mathcal{Z}$, locally asymptotically stabilizes the origin $0 \in \mathbb{R}^3$ and $V(z_{\mathrm{R}}) = \frac{1}{2}(r^2 + \frac{k_\beta}{k_r}(\alpha_{\mathrm{R}} - \beta_{\mathrm{R}})^2 + \alpha_{\mathrm{R}}^2)$ is monotonically decreasing for all $z_{\mathrm{R}} \in \mathcal{Z}$. ⌟

Proof. First note that $\lim_{\alpha_{\mathrm{R}}\to 0}\frac{\sin(\alpha_{\mathrm{R}})}{\alpha_{\mathrm{R}}} = 1$ and thus the feedback law is well-defined. Moreover, since the matrix $\begin{bmatrix} k_\beta + k_r & -k_\beta \\ -k_\beta & k_\beta \end{bmatrix}$ is positive definite, V is radially unbounded. Extending the derivations in Lemma 2, it holds that

$$\begin{aligned}&\langle \nabla V(z_{\mathrm{R}}), f_{\mathrm{R}}(z_{\mathrm{R}}, v_{\mathrm{R}})\rangle \\ &= r^2 v_{\mathrm{R}_1}\cos(\alpha_{\mathrm{R}}) + \tfrac{k_\beta}{k_r}(-v_{\mathrm{R}_2} - v_{\mathrm{R}_1}\sin(\alpha_{\mathrm{R}}) + v_{\mathrm{R}_2})(\alpha_{\mathrm{R}} - \beta_{\mathrm{R}}) \\ &\quad + \alpha_{\mathrm{R}}(-v_{\mathrm{R}_2} - v_{\mathrm{R}_1}\sin(\alpha_{\mathrm{R}})) \\ &= -k_r r^2\cos^2(\alpha_{\mathrm{R}}) - k_\alpha\alpha_{\mathrm{R}}^2 + k_\beta(\alpha_{\mathrm{R}} - \beta_{\mathrm{R}})\cos(\alpha_{\mathrm{R}})\sin(\alpha_{\mathrm{R}}) \\ &\quad - \alpha_{\mathrm{R}} k_\beta(\alpha_{\mathrm{R}} - \beta_{\mathrm{R}})\left(\cos(\alpha_{\mathrm{R}})\tfrac{\sin(\alpha_{\mathrm{R}})}{\alpha_{\mathrm{R}}}\right) \\ &= -k_r r^2\cos^2(\alpha_{\mathrm{R}}) - k_\alpha\alpha_{\mathrm{R}}^2 \le 0 \quad \forall\, z_{\mathrm{R}} \in \mathcal{Z}\end{aligned}$$

and thus, local stability of the origin follows. Moreover, for all $z_{\mathrm{R}} \in \mathcal{Z}$ for which V is not strictly decreasing it holds that $\dot{z}_{\mathrm{R}}^T = [\,0\ k_\beta\beta_{\mathrm{R}}\ k_\beta\beta_{\mathrm{R}}\,]$, whose right-hand side is unequal to zero for all $\beta_{\mathrm{R}} \neq 0$. Hence, local asymptotic stability follows from the Krasovskii-LaSalle invariance theorem (Vidyasagar, 1993, Theorem 5.3.77). □

Observe that through $k_\beta = 0$, Lemma 3 covers the result of Lemma 2 as a special case because (13) reduces to (12). For the z_{F}-dynamics in (6), using the same ideas, a result equivalent to Lemma 3 can be derived. We summarize this result in the following corollary.

Corollary 1. Let $k_r, k_\alpha, k_\beta \in \mathbb{R}_{>0}$ be arbitrary. Consider the dynamics (6) with $z_{\mathrm{F}} \in \mathcal{Z}$. Then the control law

$$v_{\mathrm{F}} = \begin{bmatrix} k_r\cos(\alpha_{\mathrm{F}}) \\ k_r\cos(\alpha_{\mathrm{F}})\sin(\alpha_{\mathrm{F}}) + k_\alpha\alpha_{\mathrm{F}} + k_\beta(\alpha_{\mathrm{F}} - \beta_{\mathrm{F}})\frac{\cos(\alpha_{\mathrm{F}})\sin(\alpha_{\mathrm{F}})}{\alpha_{\mathrm{F}}} \end{bmatrix} \tag{14}$$

is well-defined for all $z_{\mathrm{F}} \in \mathcal{Z}$, locally asymptotically stabilizes the origin $0 \in \mathbb{R}^3$ and $V(z_{\mathrm{F}}) = \frac{1}{2}(r^2 + \frac{k_\beta}{k_r}(\alpha_{\mathrm{F}} - \beta_{\mathrm{F}})^2 + \alpha_{\mathrm{F}}^2)$ is monotonically decreasing for all $z_{\mathrm{F}} \in \mathcal{Z}$. ⌟

4. A GLOBAL HYBRID STABILIZER

In this section we combine the two local control laws introduced in the preceding section in a hybrid systems formulation. As a first step, we introduce an additional discrete variable $q \in \{-1, 1\}$ where $q = -1$ represents the rear camera R and $q = 1$ represents the front camera F. In the overall system representation we consider the state

$$\xi = [r\ \beta\ \alpha\ q]^T \in \Xi \tag{15a}$$

where the pair (α, β) can either represent $(\alpha_{\mathrm{R}}, \beta_{\mathrm{R}})$ when $q = -1$ or $(\alpha_{\mathrm{F}}, \beta_{\mathrm{F}})$ when $q = 1$ and where the domain Ξ is defined as

$$\Xi := \mathbb{R}_{\ge 0} \times \mathbb{R} \times [-\tfrac{\pi}{2} - \delta, \tfrac{\pi}{2} + \delta] \times \{-1, 1\} \tag{15b}$$

for $\delta \in (0, \frac{\pi}{2})$. With this definition, the dynamics (6) can be summarized through the flow map

$$\dot{\xi} = \begin{bmatrix} \dot{r} \\ \dot{\beta} \\ \dot{\alpha} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} -qv_1 r\cos(\alpha) \\ -v_2 \\ -v_2 + qv_1\sin(\alpha) \\ 0 \end{bmatrix}, \quad \xi \in \mathcal{C} \tag{15c}$$

and the feedback laws (13) and (14) are captured through

$$v = \begin{bmatrix} qk_r\cos(\alpha) \\ k_r\cos(\alpha)\sin(\alpha) + k_\alpha\alpha + k_\beta(\alpha - \beta)\cos(\alpha)\frac{\sin(\alpha)}{\alpha} \end{bmatrix} \tag{15d}$$

The set $\mathcal{C}$, denoting the flow set, is defined as

$$\mathcal{C} := \{\xi \in \Xi : |\alpha| \le \tfrac{\pi}{2} + \delta, |\beta| \le \tfrac{3}{2}\pi + \delta\}.$$

For the jump map and the jump set, we first define the functions

$$g_\beta(\xi) = \begin{bmatrix} r \\ \beta - 2\pi\,\mathrm{sign}(\beta) \\ \alpha \\ q \end{bmatrix},\ g_\alpha(\xi) = \begin{bmatrix} r \\ \beta \\ \alpha - \pi\,\mathrm{sign}(\alpha) \\ -q \end{bmatrix}$$

and the sets

$$\begin{aligned}\mathcal{D}_\alpha &:= \{\xi \in \Xi : |\alpha| \ge \tfrac{\pi}{2} + \delta \ \wedge\ |\beta| \le \tfrac{3}{2}\pi + \delta\}, \\ \mathcal{D}_\beta &:= \{\xi \in \Xi : |\beta| \ge \tfrac{3}{2}\pi + \delta\}.\end{aligned}$$

Then the jump map is defined as

$$\xi^+ \in G(\xi) = \begin{cases} \{g_\beta(\xi)\} & \text{if } \xi \in \mathcal{D}_\beta \backslash \mathcal{D}_\alpha, \\ \{g_\alpha(\xi)\} & \text{if } \xi \in \mathcal{D}_\alpha \backslash \mathcal{D}_\beta, \\ \{g_\alpha(\xi)\} \cup \{g_\beta(\xi)\} & \text{if } \xi \in \mathcal{D}_\alpha \cap \mathcal{D}_\beta, \end{cases} \tag{15e}$$

and the jump set is defined as the union

$$\mathcal{D} := \mathcal{D}_\beta \cup \mathcal{D}_\alpha. \tag{15f}$$

Note that $\beta^+ = \beta - 2\pi\,\mathrm{sign}(\beta)$ defined through g_β guarantees that β^+ and β differ by a multiple of 2π and $|\beta^+| < |\beta|$ for all $\xi \in \mathcal{D}_\beta$ (wherein $|\beta| \ge \frac{3}{2}\pi + \delta$). Thus, β^+ and β describe the same information with respect to the position of the robot but β^+ is closer to the target orientation $\beta = 0$. Similarly, $\alpha^+ = \alpha - \pi\,\mathrm{sign}(\alpha)$ defined through g_α captures the properties in (5) when the perspective of the cameras is switched. Additionally, from the definition of the hybrid system it is clear that multiple consecutive jumps are possible, but, due to the selection of the parameter δ, Zeno behavior is not possible. Finally, since (15) satisfies (Goebel et al., 2012, As. 6.5), then asymptotic stability is robust in the sense of (Goebel et al., 2012, Ch. 7).

Theorem 1. Let $\delta \in (0, \frac{\pi}{2})$, $k_r, k_\alpha \in \mathbb{R}_{>0}$ and $k_\beta \in (0, \frac{2\delta k_r}{3\pi}]$ be arbitrary. Then, the set $\mathcal{A}_q = \{0\} \times \{0\} \times \{0\} \times \{-1, 1\}$ is globally robustly asymptotically stable for the hybrid closed-loop system dynamics (15). Moreover, $V(\xi) = \frac{1}{2}(r^2 + \frac{k_\beta}{k_r}(\alpha - \beta)^2 + \alpha^2)$ is monotonically decreasing along solutions $\xi : \mathrm{dom}(\xi) \to \Xi$. ⌟

Proof. We have established local properties of the closed-loop dynamics in Lemma 3 and in Corollary 1. What is left to show, is that the function V is decreasing at discrete time updates. Let $\xi \in \mathcal{D}_\beta$. Then it holds that

$$\begin{aligned}V(\xi^+) - V(\xi) &= \tfrac{k_\beta}{2k_r}(\alpha^+ - \beta^+)^2 - \tfrac{k_\beta}{2k_r}(\alpha - \beta)^2 \\ &= \tfrac{k_\beta}{2k_r}\left[4(\alpha - \beta)\,\mathrm{sign}(\beta)\pi + 4\pi^2\right] \\ &= 2\tfrac{k_\beta}{k_r}\left[\alpha\,\mathrm{sign}(\beta)\pi - \beta\,\mathrm{sign}(\beta)\pi + \pi^2\right]\end{aligned}$$

$$\leq 2\tfrac{k_\beta}{k_r}\left[(\tfrac{\pi}{2}+\delta)\pi-(\tfrac{3}{2}\pi+\delta)\pi+\pi^2\right]=0.$$

Similarly, for $\xi \in \mathcal{D}_\alpha$ it holds that

$$\begin{aligned}
V(\xi^+)-V(\xi) &= \tfrac{k_\beta}{2k_r}(\alpha^+-\beta^+)^2+\tfrac{1}{2}(\alpha^+)^2-\tfrac{k_\beta}{2k_r}(\alpha-\beta)^2-\tfrac{1}{2}\alpha^2\\
&= \tfrac{k_\beta}{k_r}(\alpha-\beta)\operatorname{sign}(\alpha)\pi+\tfrac{k_\beta}{2k_r}\pi^2-|\alpha|\pi+\tfrac{1}{2}\pi^2\\
&= -\tfrac{k_\beta}{k_r}|\alpha|\pi+\tfrac{k_\beta}{k_r}\beta\operatorname{sign}(\alpha)\pi+\tfrac{k_\beta}{2k_r}\pi^2-|\alpha|\pi+\tfrac{1}{2}\pi^2\\
&= -\tfrac{k_\beta}{k_r}|\alpha|\pi+\beta\operatorname{sign}(\alpha)\pi+\tfrac{k_\beta}{2k_r}\pi^2-|\alpha|\pi+\tfrac{1}{2}\pi^2\\
&\leq -\tfrac{k_\beta}{k_r}(\tfrac{\pi}{2}+\delta)\pi+\tfrac{k_\beta}{k_r}(\tfrac{3\pi}{2}+\delta)\pi+\tfrac{k_\beta}{2k_r}\pi^2-(\tfrac{\pi}{2}+\delta)\pi+\tfrac{1}{2}\pi^2\\
&\leq -\delta\pi+\tfrac{3}{2}\tfrac{k_\beta}{k_r}\pi^2\leq 0
\end{aligned}$$

and where the last inequality follows from the assumption $k_\beta \leq \frac{2\delta k_r}{3\pi}$. Thus, V is monotonically decreasing and we can conclude global asymptotic stability. Finally, robustness follows from (Goebel et al., 2012, Thm 7.21). □

5. NUMERICAL SIMULATIONS

We illustrate the results derived in the preceding section based on numerical simulations. Figure 3 shows closed-loop solutions using the feedback law (15d) with $k_\beta = 0$, i.e., the final orientation ϕ (or β) is not penalized. In particular, closed-loop solutions for various initial conditions in the (p_1, p_2)-plane, as well as the evolution over time in the x and z coordinates are shown. The remaining gains are defined as $k_r = 2$ and $k_\alpha = 1$, respectively. Additionally, the parameter $\delta = \frac{\pi}{10}$ is used for the simulations. To illustrate robustness properties of the controller, ξ is replaced by $\xi + [\varepsilon_r\ \varepsilon_\beta\ \varepsilon_\alpha\ 0]^T$ in the right-hand side of (15c) in the simulations, where ε_r, ε_β and ε_α represent white Gaussian noise with zero mean and standard deviations $\sigma_r = 0.05$, $\sigma_\beta = \sigma_\alpha = \frac{3\pi}{180}$. As expected from the theoretical results, r and α converge to to origin, while the angle β does not necessarily converge to zero.

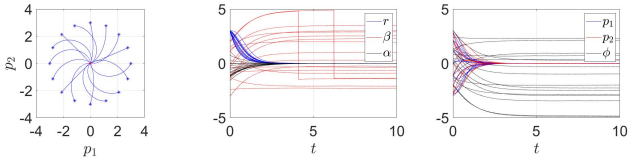


Fig. 3. Closed-loop solutions of the hybrid system (15) with controller gain $k_\beta = 0$.

For the simulations in Figure 4 the gain $k_\beta = 0$ has been replaced by $k_\beta = \frac{2\delta k_r}{3\pi}$. As expected, the controller ensures that additionally the orientation in terms of β or ϕ, respectively converges to zero for $t \to \infty$ according to Theorem 1. Figure 4 additionally shows the decrease of the function V defined in Theorem 1.

6. CONCLUSIONS

Inspired by the controller design in Aicardi et al. (1995), in this work we have proposed a globally stabilizing controller for unicycle dynamics relying on a hybrid systems formulation. The controller is motivated through mobile robots equipped with range sensors and front and rear cameras with overlapping fields of view.

While the control law derived in Theorem 1 is unbounded, a bounded globally stabilizing control law can be obtained by appropriately scaling v in (15d) (see (Braun et al., 2021, Theorem 2.3), for example). Such a scaling

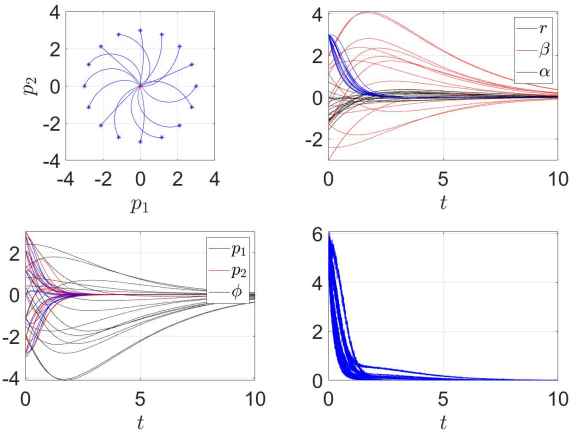


Fig. 4. Closed loop solutions of the hybrid system (15).

can also be used to handle unknown input gains, which have been encountered for example in Vinco et al. (2021), where the input gain depends on the (unknown) state of charge of the battery. Future work will focus on the analysis of robustness properties of the controller and will incorporate obstacle avoidance properties in the overall controller design. In this context we will take inspiration from Braun and Zaccarian (2021) and Marley et al. (2021).

REFERENCES

Aicardi, M., Casalino, G., Bicchi, A., and Balestrino, A. (1995). Closed loop steering of unicycle like vehicles via lyapunov techniques. *IEEE Robotics Automation Magazine*, 2(1), 27–35.

Braun, P., Grüne, L., and Kellett, C.M. (2021). *(In-)Stability of Differential Inclusions: Notions, Equivalences, and Lyapunov-like Characterizations*. Springer.

Braun, P. and Zaccarian, L. (2021). Augmented obstacle avoidance controller design for mobile robots. *IFAC-PapersOnLine*, 54(5), 157–162. 7th IFAC Conference on Analysis and Design of Hybrid Systems.

Brockett, R.W. (1983). Asymptotic stability and feedback stabilization. *Differential Geometric Control Theory*, 27(1), 181–191.

Goebel, R., Sanfelice, R.G., and Teel, A.R. (2012). *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*. Princeton University Press.

Marley, M., Skjetne, R., and Teel, A.R. (2021). Synergistic control barrier functions with application to obstacle avoidance for nonholonomic vehicles. In *2021 American Control Conference*, 243–249.

Nešić, D., Teel, A., and Zaccarian, L. (2011). Stability and performance of SISO control systems with First Order Reset Elements. *IEEE Trans. on Automatic Control*, 56(11), 2567–2582.

Sontag, E.D. (1999). Nonlinear feedback stabilization revisited. In *Dynamical Systems, Control, Coding, Computer Vision*, 223–262. Birkhäuser Basel.

Tzafestas, S.G. (2013). *Introduction to Mobile Robot Control*. Elsevier.

Vidyasagar, M. (1993). *Nonlinear Systems Analysis: Second Edition*. Prentice-Hall.

Vinco, G.M., Braun, P., and Zaccarian, L. (2021). A modular architecture for mobile robots equipped with continuous-discrete observers. In *IEEE International Conference on Mechatronics*, 1–6.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Kernelized Active Subspaces

Benjamin P. Russo * **Joel A. Rosenfeld** **

* *Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831,* russobp@ornl.gov
** *Department of Mathematics and Statistics University of South Florida Tampa, Fl 33620,* rosenfeldj@usf.edu

Abstract: Given a C^1 function over a potentially high dimensional domain, the active subspace method seeks an affine subspace inside which the functions changes the most on average. This is done by finding the eigenvectors of a covariance matrix incorporating gradient information. In a similar vein, the active manifold method finds a manifold γ and if information on f is recovered along γ then f can be recovered on the connected component of a level set touching γ. An inherent limitation of the Active subspace technique is that it only considers affine subspaces (which may still be high dimensional). Inspired by methods in occupation kernel dynamic mode decomposition, we develop a notion of active subspace taking place in a Hilbert space which contains sufficient complexity to describe highly nonlinear level sets. In this learning problem, only function values along trajectories following the gradient direction of the function are required to determine this decomposition.

Keywords: active manifolds, occupation kernels, Liouville operators, reproducing kernels.

1. INTRODUCTION

Active subspaces is a dimension reduction technique originated by (Russi (2010)) and developed extensively by (Constantine (2015)). Current implementations of the active subspace method can capture "active" subspaces for functions whose level sets are subspaces of $\mathbb{R}^n$. However, for functions that have very nonlinear levelsets, their applicability is limited to a small neighborhood, where the functions may be effectively linearized. This limitation manifests theoretically in the covariance matrix leveraged in the construction of active subspaces which is finite dimensional and lacks the complexity necessary for discovering level-sets that are nonlinear manifolds.

To address this limitation, the present manuscript introduces a new operator valued kernel to stand in place of the covariance matrix of Constantine (2015). This operator valued kernel, based on the Liouville operator given in Section 4, acts on a vector valued RKHS, which is infinite dimensional. This kernel makes available an infinite collection of eigenvalues and eigenvectors that can be leveraged to decompose a function into "active" and "inactive" components, where the active component will effectively represent level sets of the original function. In this learning problem, only function values along trajectories following the gradient direction of the function are required to determine this decomposition. Significantly, the gradient of the original function need not be computed using the present method, which differs from established work (cf. Bridges et al. (2019)).

2. REVIEW OF ACTIVE SUBSPACE METHODS

Let $X \subseteq \mathbb{R}^m$ be a compact subset equipped with a probability measure ρ and suppose that $f : X \to \mathbb{R}$ is a $C^1(X)$ function. Let, $\nabla_x f = [\partial f/\partial x_1, \ldots, \partial f/\partial x_n]^\top$ and define the $n \times n$ matrix C given by

$$C = \mathbb{E}[(\nabla_x f)(\nabla_x f)^\top] = \left(\int_X \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \, d\rho \right)_{i,j=1,1}^{n,n}. \quad (1)$$

Since C is a positive semidefinite square matrix it admits an eigenvalue decomposition

$$C = W \Lambda W^\top, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_1 \geq \ldots \geq \lambda_n \geq 0.$$

For a given $m < n$ we can further decompose Λ into a block diagonal matrix $\Lambda = \operatorname{diag}(\Lambda_1, \Lambda_2)$ where $\Lambda_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ and $\Lambda_2 = \operatorname{diag}(\lambda_{m+1}, \ldots, \lambda_n)$. Likewise W can be decomposed as $W = [W_1 \, W_2]$, where W_1 is $n \times m$, and

$$C = [W_1 \; W_2] \operatorname{diag}(\Lambda_1, \Lambda_2) [W_1 \; W_2]^\top.$$

Let w_i^1 for $i = 1, \ldots, m$ be the column vectors defined by W_1, For our choice of m we define $\mathcal{A} = \operatorname{span}\{w_i^1 \mid i = 1, \ldots, m\}$ as the active subspace of f. We can interpret C as the uncentered covariance for the gradient and since the eigenvalues in the decomposition are listed in decreasing order the active subspace A corresponds to directions which have greater average variation for the function f. In this paper we develop a generalized version of the active subspace method that enables the learning of nonlinear level sets through the introduction of an operator valued kernel.

3. MATHEMATICAL PRELIMINARIES

The main setting for this generalized approach is vector valued reproducing kernel Hilbert spaces (RKHS). For completeness we will also define reproducing kernel Hilbert spaces.

Definition 3.1. A RKHS, H, over a set X is a Hilbert space of real valued functions over the set X such that for all $x \in X$ the evaluation functional $E_x g := g(x)$ is bounded.

As such, the Riesz representation theorem guarantees, for all $x \in X$, the existence of a function $k_x \in H$ such that $\langle g, k_x\rangle_H = g(x)$, where $\langle\cdot,\cdot\rangle_H$ is the inner product for H.

The function k_x is called the reproducing kernel function at x, and the function $k(x,y) = \langle k_y, k_x\rangle_H$ is called the kernel function corresponding to H.

Vector valued RKHSs began to appear in learning theory over the past decade Carmeli et al. (2010), though their inception dates back at least as far as the 1950s (e.g. Pedrick (1957)).

Definition 3.2. Given a Hilbert space $\mathcal{Y}$ and a set X, a vector valued reproducing kernel Hilbert space, H, is a Hilbert space of functions mapping X to $\mathcal{Y}$, where for each $x \in X$ the evaluation mapping $E_x : H \to \mathcal{Y}$ given by $E_x(f) = f(x)$ is bounded. The operator valued kernel for a vector valued reproducing kernel Hilbert space is given by $K : X \times X \to \mathcal{B}(Y)$, $K(x,y) = E_x E_y^*$, here $\mathcal{B}(\mathcal{Y})$ denotes the bounded operators on $\mathcal{Y}$.

Boundedness of the functional $E_x : H \to \mathcal{Y}$ is equivalent to the boundedness of the functional $H \ni g \mapsto \langle g(x), y\rangle_{\mathcal{Y}}$ for each $x \in X$ and $y \in \mathcal{Y}$. The Riesz representation theorem guarantees for each $x \in X$ and $v \in \mathcal{Y}$ the existence of a function $K_{x,v} \in H$ such that $\langle g, K_{x,v}\rangle_H = \langle g(x), v\rangle_{\mathcal{Y}}$ for all $g \in H$.

Remark 1. In general, for a vector valued reproducing kernel Hilbert space H with kernel K, we can define $K_x v(\cdot) := K(\cdot, x)v \in H$. We note that $K_{x,v} = K_x v = E_x^*(v)$ since

$$(E_x^*(v))(y) = E_y E_x^*(v) = K(y,x)v = (K_x v)(y)$$

for all y and the reproducing property is given by

$$\langle f(x), v\rangle_{\mathcal{Y}} = \langle E_x(y), v\rangle_{\mathcal{Y}} = \langle f, E_x^*(v)\rangle_H = \langle f, K_{x,v}\rangle_H.$$

Definition 3.3. Let $X \subset \mathbb{R}^n$ be compact, H be a $\mathbb{R}^n$ valued RKHS of continuous functions over X, and $\gamma : [0,T] \to X$ be a bounded measurable trajectory. For every $v \in \mathbb{R}^n$, the functional $g \mapsto \left\langle \int_0^T g(\gamma(\tau))d\tau, v\right\rangle_{\mathbb{R}^n}$ is bounded, and may be respresented as $\left\langle \int_0^T g(\gamma(\tau))d\tau, v\right\rangle_{\mathbb{R}^n} = \langle g, \Gamma_{\gamma,v}\rangle_H$, for some $\Gamma_{\gamma,v} \in H$ by the Riesz representation theorem. The function $\Gamma_{\gamma,v}$ is called the occupation kernel corresponding to γ in H and $v \in \mathbb{R}^n$

Definition 3.4. Let $X \subset \mathbb{R}^n$ be compact, H be a $\mathcal{Y}$-valued RKHS of continuous functions over X, and $\gamma : [0,T] \to X$ be a bounded measurable trajectory. Define the operators E_γ and E_γ^* by

$$E_\gamma : H \to \mathcal{Y}, \quad g \mapsto \int_0^T g(\gamma(t))dt \in \mathcal{Y}$$

and

$$E_\gamma^* : \mathcal{Y} \to H, \quad v \mapsto \Gamma_{\gamma,v}.$$

We denote $E_\gamma^*(v) = \Gamma_{\gamma,v}$ by $\Gamma_\gamma v$. Let P be the set of bounded measurable trajectories. We call

$$\Gamma(x,\gamma) : X \times P \to \mathcal{B}(\mathcal{Y}), \quad \Gamma(x,\gamma) := E_x E_\gamma^*$$

the operator valued occupation kernel.

Proposition 3.5. For all $i \in \{1,\dots,n\}$ let $\mathscr{H}_i$ be real-valued reproducing kernel Hilbert spaces on a set X and $H = \bigoplus_{i=1}^n \mathscr{H}_i$ be the associated $\mathbb{R}^n$-valued reproducing kernel Hilbert space. For a given path $\gamma : [0,T] \to X \subset \mathbb{R}^k$, let $\Gamma_\gamma^i \in \mathscr{H}_i$ be the scalar valued occupation kernel associated to γ. The function $\Gamma_{\gamma,v} \in H$ is given by

$$\Gamma_\gamma(x) \odot v$$

where $\Gamma_\gamma(x) = (\Gamma_\gamma^1(x), \dots, \Gamma_\gamma^n(x))^\top$. Here, $\odot$ represents the Hadamard product.

Proof. Let $g = (g_1, \dots, g_n)^\top$ be in H, and $v = (v_1, \dots, v_n)^\top \in \mathbb{R}^n$ then

$$\begin{aligned}\langle g, \Gamma_{\gamma,v}\rangle_H &= \left\langle \int_0^T g(\gamma(t))dt, v\right\rangle_{\mathbb{R}^n} \\ &= \sum_{i=1}^n \int_0^T v_i g_i(\gamma(t))dt \\ &= \sum_{i=1}^n \langle g_i, \Gamma_\gamma^i v_i\rangle_{\mathscr{H}_i} = \langle g, \Gamma_\gamma \odot v\rangle_H.\end{aligned}$$

4. THEORETICAL DESCRIPTION OF THE METHOD

For this section we will assume the functions are over a compact domain $X \subset \mathbb{R}^n$ and are $\mathbb{R}^n$ valued.

Definition 4.1. For an $\mathbb{R}^n$ valued reproducing kernel Hilbert space H and a given symbol $\psi : \mathbb{R}^n \to \mathbb{R}^n$ we define the Liouville operator with symbol ψ as $A_\psi(g) = D(g)\psi$.

Modally, A_ψ is a closed, densely defined, and unbounded operator, owing to the inclusion of the differentiation operator D. Throughout this manuscript, a heuristic assumption that this operator is bounded and even compact will be leveraged in the development of the numerical methods of this paper. This heuristic assumption is justified in several contexts, where the selection of appropriate Hilbert spaces in the range and domain of A_ψ can make it a bounded operator, and the use of scaled versions of this operator (cf. Rosenfeld et al. (2021)) can in fact produce a compact operator that agrees to computational precision with A_ψ on a compact subset of a given workspace.

Given a $f, h \in C^1(\mathbb{R}^n, \mathbb{R})$, define the Liouville operator $A_{\nabla f} : \mathcal{D}(A_{\nabla f}) \to H$. Let $\mathcal{D}_h(A_{\nabla f}) \subset \mathcal{D}(A_{\nabla f})$ be defined as those vectors $g \in \mathcal{D}(A_{\nabla f})$ such that $A_{\nabla f} g \in \mathcal{D}(A_{\nabla h}^*)$. Under the heuristic assumption discussed above, $\mathcal{D}_h(A_{\nabla f}) = H$. Definition 4.2 yields an operator theoretic replacement for the covariance matrix. The advantage gained through this perspective is that the operator valued kernel in Definition 4.2 provides a potentially infinite collection of eigenvalues and eigenfunctions that can be leveraged to decompose $\mathbb{R}^n$ in a nonlinear manner.

Definition 4.2. Let H be a $\mathbb{R}^n$- valued reproducing kernel Hilbert space, we define

$$C(\nabla f, \nabla h) : \mathcal{D}_h(A_f) \to H, \qquad C(\nabla f, \nabla h) := A_{\nabla h}^* A_{\nabla f}.$$

If it is assumed that $A_{\nabla f}$ is compact, which may require an adjustment between the domain and range of the operator (cf. Rosenfeld and Kamalapurkar (2021)), and consequently $C(\nabla f, \nabla f)$ is compact (compact operators form an ideal in the algebra of bounded operators), then as a self adjoint operator, $C(\nabla f, \nabla f)$ is diagonalizable. That is, the eigenfunctions of $C(\nabla f, \nabla f)$ form an orthonormal basis of H. Moreover, $A_{\nabla f}$ has a singular value decomposition, where the right singular vectors are the eigenfunctions of

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$C(\nabla f, \nabla f)$. In the subsequent numerical methods, finite rank representations for $C(\nabla f, \nabla f)$ will be extracted from finite rank representations of $A_{\nabla f}$.

Definition 4.3. Suppose $\varepsilon > 0$ is a given threshold and that $A_{\nabla f}$ is diagonalizable. Define,

$$\mathcal{A}_\varepsilon := \text{span}\{\varphi \mid C(\nabla f, \nabla f)\varphi = \lambda\varphi, \quad \sqrt{|\lambda|} > \varepsilon\}$$
$$\mathcal{I}_\varepsilon := \mathcal{A}_\varepsilon^\perp.$$

Note, our Hilbert space H can be orthogonally decomposed as $H = \mathcal{A}_\varepsilon \oplus \mathcal{I}_\varepsilon$. Furthermore, for notational convenience write $\Sigma := \{\varphi \in H : \exists \lambda \in \mathbb{C}$ such that $C(\nabla f, \nabla f)\varphi = \lambda\varphi\}$ and $\Sigma^* := \{\psi \in H : \exists \lambda \in \mathbb{C}$ such that $A_{\nabla f}A_{\nabla f}^*\psi = \lambda\psi\}$. Each $\varphi \in \Sigma$ is a right singular vector of $A_{\nabla f}$ and maps to the left singular vectors under$A_{\nabla f}$. Moreover, the identity function, $g_{id}(x) = x$, which is assumed to be in H, admits the decomposition

$$x = g_{id}(x) = \sum_{\varphi \in \Sigma} \langle g, \varphi \rangle_H \varphi(x).$$

Set $M = \max_{\varphi \in \Sigma} |\langle \varphi, g_{id} \rangle_H|$ and set $\varepsilon_0 = \varepsilon / M$. Write $\Sigma_\varepsilon := \{\varphi \in \Sigma : \sqrt{|\lambda|} > \varepsilon\}$. Then,

$$\begin{aligned}
&\nabla f(x) \\
&= D(g_{id})\nabla f(g_{\text{id}}(x)) \\
&= D(P_{\mathcal{A}_{\varepsilon_0}} g_{\text{id}}(x))\nabla f(x) + D(P_{\mathcal{I}_{\varepsilon_0}} g_{\text{id}}(x))\nabla f(x) \\
&= \sum_{\varphi \in \Sigma_{\varepsilon_0}} \langle g_{id}, \varphi \rangle_H A_{\nabla f}\varphi(x) + \sum_{\varphi \in \Sigma \setminus \Sigma_{\varepsilon_0}} \langle g_{id}, \varphi \rangle_H A_{\nabla f}\varphi(x) \\
&= \sum_{\varphi \in \Sigma_{\varepsilon_0}} \langle g_{id}, \varphi \rangle_H \sqrt{\lambda}\psi(x) + \sum_{\varphi \in \Sigma \setminus \Sigma_{\varepsilon_0}} \langle g_{id}, \varphi \rangle_H \sqrt{\lambda}\psi(x),
\end{aligned}$$

where ψ are the right singular vectors of $A_{\nabla f}$ corresponding to $\varphi \in \Sigma$.

Definition 4.4. The subspace $\mathcal{A}_{\varepsilon_0}$ described above will be called the *active subspace* within H for f and $P_{\mathcal{A}_{\varepsilon_0}} g_{\text{id}}$ is called the *active* component of g_{id} (of order ε)

Each term of the active component satisfies

$$\langle g_{id}, \varphi \rangle \sqrt{|\lambda|} > \varepsilon.$$

5. FINITE RANK REPRESENTATIONS VIA OCCUPATION KERNELS

To computationally determine an estimation of the eigendecomposition of $C(\nabla f, \nabla f)$, a finite rank representation of A_f will be determined using a collection of trajectories, $\{\gamma_i : [0, T] \to \mathbb{R}^n\}_{i=1}^M$, that satisfy $\dot{\gamma}_i = \nabla f(\gamma_i)$ and their corresponding operator valued occupation kernels, $\{\Gamma_\gamma\}_{i=1}^M$. Significant to the method is the following proposition:

Proposition 5.1. Let H be a $\mathbb{R}^n$ valued RKHS consisting of continuously differentiable funcitons, and let $\gamma : [0, T] \to \mathbb{R}^n$ be a trajectory satisfying $\dot{\gamma} = \nabla f(\gamma)$ for a function $f : \mathbb{R}^n \to \mathbb{R}$ which yields a densely defined Liouville operator, $A_{\nabla f}$, over H. Then the following relation holds

$$\langle A_{\nabla f} g, \Gamma_\gamma v \rangle_H = \langle g, (K_{\gamma(T)} - K_{\gamma(0)})v \rangle_H. \tag{2}$$

Hence, $A_{\nabla f}^* \Gamma_{\gamma, v} = (K_{\gamma(T)} - K_{\gamma(0)})v$.

Proof. Note that

$$\begin{aligned}
\langle A_{\nabla f} g, \Gamma_{\gamma, v} \rangle_H &= \langle D(g)\nabla f, \Gamma_{\gamma, v} \rangle_H \\
&= \left\langle \int_0^T Dg(\gamma(t))\nabla f(\gamma(t))dt, v \right\rangle_{\mathbb{R}^n} \\
&= \left\langle \int_0^T \frac{d}{dt} g(\gamma(t))dt, v \right\rangle_{\mathbb{R}^n} \\
&= \langle g(\gamma(T)) - g(\gamma(0)), v \rangle_{\mathbb{R}^n} \\
&= \langle g, \left(K_{\gamma(T)} - K_{\gamma(0)}\right) v \rangle_H.
\end{aligned}$$

Leveraging this relation, a finite rank representation of $A_{\nabla f}^*$ may be determined, and consequently the transpose of this representation will represent $A_{\nabla f}$ under the boundedness assumption. In particular, replacing v with vectors from the standard basis in $\mathbb{R}^n$, and for a fixed $1 \leq i \leq n$ writing $\beta_i = \text{span}\{\Gamma_{\gamma_j} e_i\}_{j=1}^M$ will give us the following proposition.

Proposition 5.2. Let $\{e_k\}$ denote the standard basis for $\mathbb{R}^n$ and fix an $i \in \{1, \ldots, n\}$. For a finite dimensional subspace given by $\beta_i = \text{span}\{\Gamma_{\gamma_j, e_i}\}_{j=1}^M = \text{span}\{\Gamma_{\gamma_j} e_i\}_{j=1}^M$,

$$[P_{\beta_i} A_{\nabla f}^*]_{\beta_i}^{\beta_i} = \begin{pmatrix} \langle \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_1} e_i \rangle_H & \cdots & \langle \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_1} e_i \rangle_H \\ \vdots & \ddots & \vdots \\ \langle \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_M} e_i \rangle_H & \cdots & \langle \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_M} e_i \rangle_H \end{pmatrix}^{-1}$$
$$\times \begin{pmatrix} \langle (K_{\gamma_1(T)} - K_{\gamma_1(0)}) e_i, \Gamma_{\gamma_1} e_i \rangle_H & \cdots & \langle (K_{\gamma_M(T)} - K_{\gamma_M(0)}) e_i, \Gamma_{\gamma_1} e_i \rangle_H \\ \vdots & \ddots & \vdots \\ \langle (K_{\gamma_1(T)} - K_{\gamma_1(0)}) e_i, \Gamma_{\gamma_M} e_i \rangle_H & \cdots & \langle (K_{\gamma_M(T)} - K_{\gamma_M(0)}) e_i, \Gamma_{\gamma_M} e_i \rangle_H \end{pmatrix}.$$

Proof. For $h \in \mathcal{D}(A_{\nabla f}^*)$, the coefficients $\{a_j\}_{j=1}^M$ in the projection of $A_{\nabla f}^* h$ onto β_i, given by $P_{\beta_i} A_{\nabla f}^* h = \sum_{j=1}^M a_j \Gamma_{\gamma_j} e_i$, can be expressed as

$$\begin{pmatrix} a_1 \\ \vdots \\ a_M \end{pmatrix} = \begin{pmatrix} \langle \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_1} e_i \rangle_H & \cdots & \langle \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_1} e_i \rangle_H \\ \vdots & \ddots & \vdots \\ \langle \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_M} e_i \rangle_H & \cdots & \langle \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_M} e_i \rangle_H \end{pmatrix}^{-1}$$
$$\times \begin{pmatrix} \left\langle A_{\nabla f}^* h, \Gamma_{\gamma_1} e_i \right\rangle_H \\ \vdots \\ \left\langle A_{\nabla f}^* h, \Gamma_{\gamma_M} e_i \right\rangle_H \end{pmatrix}.$$

Assuming that the occupation kernels are in the domain of the Liouville operator, i.e., $\beta_i \subset \mathcal{D}(A_{\nabla f}^*)$, for $h \in \beta_i$, given by $h = \sum_{j=1}^M c_j \Gamma_{\gamma_j} e_i$, for a fixed $k \in \{1, \ldots, M\}$, we have

$$\begin{aligned}
\left\langle A_{\nabla f}^* h, \Gamma_{\gamma_k} e_i \right\rangle_H &= \sum_{j=1}^M c_j \left\langle A_{\nabla f}^* \Gamma_{\gamma_j} e_i, \Gamma_{\gamma_k} e_i \right\rangle_H \\
&= \left(\left\langle A_{\nabla f}^* \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_k} e_i \right\rangle_H \ldots, \left\langle A_{\nabla f}^* \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_k} e_i \right\rangle_H \right) \\
&\quad \cdot (c_1 \ \ldots \ c_M)^\top.
\end{aligned}$$

As a result, a finite rank representation of $A_{\nabla f}^*$ restricted to β_i, i.e., the matrix $[P_{\beta_i} A_{\nabla f}^*]_\alpha^\alpha$ that maps the coefficients $\{c_j\}_{j=1}^M$ to the coefficients $\{a_j\}_{j=1}^M$, is given as

$$[P_{\beta_i} A_{\nabla f}^*]_{\beta_i}^{\beta_i} = \begin{pmatrix} \langle \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_1} e_i \rangle_H & \cdots & \langle \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_1} e_i \rangle_H \\ \vdots & \ddots & \vdots \\ \langle \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_M} e_i \rangle_H & \cdots & \langle \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_M} e_i \rangle_H \end{pmatrix}^{-1}$$
$$\times \begin{pmatrix} \langle A_{\nabla f}^* \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_1} e_i \rangle_H & \cdots & \langle A_{\nabla f}^* \Gamma_{\gamma_1} e_i, \Gamma_{\gamma_M} e_i \rangle_H \\ \vdots & \ddots & \vdots \\ \langle A_{\nabla f}^* \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_1} e_i \rangle_H & \cdots & A_{\nabla f}^* \Gamma_{\gamma_M} e_i, \Gamma_{\gamma_M} e_i \rangle_H \end{pmatrix}$$

The proof is completed by an application of Proposition 5.1

Proposition 5.3. Let $\{e_k\}$ denote the standard basis for $\mathbb{R}^n$ and for each $i \in \{1, \ldots, n\}$ let $\beta_i = \text{span}\{\Gamma_{\gamma_j, e_i}\}_{j=1}^M = \text{span}\{\Gamma_{\gamma_j} e_i\}_{j=1}^M$. Let $B = \oplus_{i=1}^n \beta_i$ then

$$[P_B A^*_{\nabla f}]_B^B = \text{diag}([P_{\beta_i} A^*_{\nabla f}]_{\beta_i}^{\beta_i})$$

Proof. Let ℓ and m be arbitrary indices in $\{1, \ldots, M\}$. We must show for $i, j \in \{1, \ldots, n\}$ with $i \neq j$ that

$$\langle A^*_{\nabla f} \Gamma_{\gamma_\ell} e_i, \Gamma_{\gamma_m} e_j \rangle = 0.$$

By an applications of Propositions 3.5 and 5.1 we get

$$\begin{aligned}\langle A^*_{\nabla f} \Gamma_{\gamma_\ell} e_i, \Gamma_{\gamma_m} e_j \rangle &= \langle \Gamma_{\gamma_\ell} e_j, (K_{\gamma_m(T)} - K_{\gamma_m(0)}) e_i \rangle_H \\ &= \langle \Gamma_{\gamma_\ell, e_j}(\gamma_m(T)) - \Gamma_{\gamma_\ell, e_j}(\gamma_m(0)), e_i \rangle_{\mathbb{R}^n}\end{aligned}$$

where we get the difference of the *scalar* valued occupation kernel for γ_ℓ evaluated at $\gamma_m(T)$ and $\gamma_m(0)$ in the j-th spot and a 1 in the i-th spot. Hence, this is non-zero only when $i = j$.

Hence for each dimension of the workspace, a collection of approximate eigenfunctions for $A_{\nabla f}$ may be determined through the SVD of the matrix $([P_{\beta_i} A^*_{\nabla f}]_{\beta_i}^{\beta_i})^T$. For each singular vector of $([P_{\beta_i} A^*_{\nabla f}]_{\beta_i}^{\beta_i})^T$, $\eta = (\eta_1, \ldots, \eta_M)$, with singular value λ, the corresponding normalized singular function in H is given as $\varphi = \frac{1}{\sqrt{\eta^T G \eta}} \sum_{j=1}^M \eta_j \Gamma_{\gamma_j} e_i$, where G is the Gram matrix corresponding to the basis for β_i. Hence, to evaluate $\sum_{\varphi \in \Sigma_{\varepsilon_0}} \langle g_{id}, \varphi \rangle_H \sqrt{\lambda} \psi(x)$ we must be able to compute $\langle g_{id}, \Gamma_{\gamma_j} e_i \rangle$. By definition, this is given as

$$\langle g_{id}, \Gamma_{\gamma_j} e_i \rangle = \left\langle \int_0^T g_{id}(\gamma_j(t)) dt, e_i \right\rangle_{\mathbb{R}^n} = \int_0^T \gamma_j^i(t) dt,$$

i.e. it is the integral of the i-th component of $\gamma_j : [0, T] \to X \subset \mathbb{R}^n$

6. DISCUSSION AND CONCLUSION

This manuscript presents a new dimension reduction technique as a nonlinear version of the active subspace method. To address the limitation of the active subspace routine this manuscript introduces a new operator valued kernel to stand in place of the covariance matrix of Constantine (2015). This operator valued kernel, based on the Liouville operator given in Section 4, acts on a vector valued RKHS, which is infinite dimensional. This kernel makes available an infinite collection of eigenvalues and eigenvectors that can be leveraged to decompose a function into "active" and "inactive" components, where the active component will effectively represent level sets of the original function. The authors expect the same challenges that appear in occupation kernel DMD to be present in this new technique as well. While, in principle any reproducing kernel Hilbert space can be used, choice of RKHS affects the operator theoretic properties of boundedness and compactness of $C(\nabla f, \nabla f)$ and $A_{\nabla f}$. (Russo and Rosenfeld (2022)) explores the properties of Liouville operators over the Hardy space and (Rosenfeld and Kamalapurkar (To Appear)) gives a modification of Liouville operators that allow for compactness. Moreover, it is found in dynamic mode decomposition that careful parameter selection is necessary and a poor choice of parameter can lead to bad reconstruction of the function. This in general is one of the major limitations of the technique.

7. ACKNOWLEDGEMENTS

This research was supported by the Air Force Office of Scientific Research (AFOSR) under contract numbers FA9550-20-1-0127 and FA9550-21-1-0134, and the National Science Foundation (NSF) under award 2027976. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

Notice: This manuscript has been authored, in part, by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (`http://energy.gov/downloads/doe-public-access-plan`).

REFERENCES

Bridges, R.A., Gruber, A.D., Felder, C., Verma, M., and Hoff, C. (2019). Active manifolds: A non-linear analogue to active subspaces.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá., V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Anal. Appl.*, 08(01), 19–61.

Constantine, P.G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies.* SIAM.

Pedrick, G. (1957). *Theory of reproducing kernels for Hilbert spaces of vector valued functions.* Ph.D. thesis, University of Kansas.

Rosenfeld, J.A. and Kamalapurkar, R. (To Appear). Dynamic mode decomposition with control liouville operators. In *Proceedings of the 24th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2021).*

Rosenfeld, J.A. and Kamalapurkar, R. (2021). Singular dynamic mode decompositions. *arXiv preprint arXiv:2106.02639.*

Rosenfeld, J.A., Kamalapurkar, R., Gruss, L.F., and Johnson, T.T. (2021). Dynamic mode decomposition for continuous time systems with the liouville operator. *Journal of Nonlinear Science*, 32(1), 5. doi:10.1007/s00332-021-09746-w. URL `https://doi.org/10.1007/s00332-021-09746-w`.

Russi, T.M. (2010). *Uncertainty quantification with experimental data and complex system models.* Ph.D. thesis, UC Berkeley.

Russo, B.P. and Rosenfeld, J.A. (2022). Liouville operators over the Hardy space. *Journal of Mathematical Analysis and Applications*, 508(2), 125854. doi: https://doi.org/10.1016/j.jmaa.2021.125854. URL `https://www.sciencedirect.com/science/article/pii/S0022247X21009367`.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Theoretical Foundations for the Dynamic Mode Decomposition of High Order Dynamical Systems

Joel A. Rosenfeld * **Benjamin P. Russo** **
Rushikesh Kamalapurkar ***

* *Department of Mathematics and Statistics, University of South Florida, Tampa, Fl 33620,* rosenfeldj@usf.edu
** *Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831,* russobp@ornl.gov
*** *School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK 74078,*
rushikesh.kamalapurkar@okstate.edu

Abstract: Conventionally, data driven identification and control problems for higher order dynamical systems are solved by augmenting the system state by the derivatives of the output to formulate first order dynamical systems in higher dimensions. However, solution of the augmented problem typically requires knowledge of the full augmented state, which requires numerical differentiation of the original output, frequently resulting in noisy signals. This manuscript develops the theory necessary for a direct analysis of higher order dynamical systems using higher order Liouville operators. Fundamental to this theoretical development is the introduction of signal valued RKHSs and new operators posed over these spaces. Ultimately, it is observed that despite the added abstractions, the necessary computations are remarkably similar to that of first order DMD methods using occupation kernels.

Keywords: system identification, operator theoretic methods in systems theory, model approximation, control system analysis

1. INTRODUCTION

Data driven methods for dynamical systems have developed significantly over the past 20 years (cf. Budišić et al. (2012); Kutz et al. (2016); Proctor et al. (2016); Mauroy and Goncalves (2020); Mauroy and Mezić (2016)). Principle among them are those that leverage Koopman operators (also known as composition operators) over Hilbert function spaces to give a representation of finite dimensional discrete time dynamics as an operator over an infinite dimensional Hilbert space Budišić et al. (2012); Williams et al. (2015). When a continuous time dynamical system is forward complete, it may be discretized by fixing a time-step, $\Delta t > 0$, to yield a discrete time system. In this setting, the Koopman operator has been demonstrated as an effective tool for extracting the underlying governing principles of a dynamical system, and for providing a model for the state which performs well over short time horizons via Dynamic Mode Decompositions (DMD).

In Rosenfeld et al. (2019b), the concept of occupation kernels was introduced as functions inside of a RKHS that, given a signal $\theta : [0, T] \to \mathbb{R}^n$, represent the functional $g \mapsto \int_0^T g(\theta(t))dt$. Occupation kernels generalize the concept of an occupation measure (cf. Lasserre et al. (2008)) by changing the setting from a collection of measures to that of a Hilbert space. Thus, occupation kernels can be leveraged as a basis in a Hilbert space for function approximation and projections (cf. Rosenfeld et al. (2019b,a); Rosenfeld and Kamalapurkar (2021, to appear, see arXiv:2101.02620); Li and Rosenfeld (to appear); Russo et al. (2022)).

One limitation still present in the theory of data driven methods for dynamical systems is that of high order dynamics. Conventionally in systems theory, higher order dynamics are converted to first order systems of augmented state variables. For example, $\ddot{x} = f(x)$ can be adjusted to $z := (x\ \dot{x})^T$ with $\dot{z} = (z_2\ f(z_1))^T$. Theoretically, the augmentation is well justified, but it is computationally problematic in data driven methods. To estimate the new

⋆ This research was supported by the Air Force Office of Scientific Research (AFOSR) under contract numbers FA9550-20-1-0127 and FA9550-21-1-0134, and the National Science Foundation (NSF) under awards 2027976 and 2027999. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

state variable of z, data driven methods must compute an approximation of the first derivative of x. Numerical derivatives can be noisy, and if the order of the system exceeds 2, they are unreliable. For example, in Brunton et al. (2016), where numerical differentiation is used for parameter fitting, a considerable amount of filtering was required to get good results. The sensitivity of numerical differentiation techniques to noise motivates the development of methods that avoid numerical differentiation altogether.

This manuscript introduces the necessary theoretical components for the development of a DMD routine for second order dynamical systems that avoids the use of numerical derivatives and augmented state variables. The exposition will be focused on second order dynamical systems. However, the developed methods may be readily adapted to higher order dynamical systems. Underlying the subsequent development are vector valued Reproducing Kernel Hilbert spaces (vvRKHSs), for which the relevant theory is presented in Section 2. Using vvRKHSs as a tool, Section 4 develops a signal valued RKHS, which is a Hilbert space of functions that map d times continuously differentiable signals to a scalar valued RKHS over $[0, T]$. The signal valued RKHS framework allows for the formulation of well defined second order Liouville operators over the Hilbert space beyond the formal expression given in Section 5. Once the essential elements are established, Section 6 presents a DMD method for the modeling of a second order dynamical system, which avoids the use of numerical derivatives.

2. VECTOR VALUED REPRODUCING KERNEL HILBERT SPACES

This section presents the concept of vector valued RKHSs, which recently came to prominence with Carmeli et al. (2010), though their roots extend further back (e.g. Pedrick (1957)). In the context of this manuscript the Hilbert space $\mathcal{Y}$ will be a scalar valued RKHS, which will facilitate the description of a function space on signals.

Definition 1. Given a Hilbert space $\mathcal{Y}$ and a set X, a vector valued reproducing kernel Hilbert space, H, is a Hilbert space of functions mapping X to $\mathcal{Y}$, where for each $x \in X$ and $v \in \mathcal{Y}$ the functional $g \mapsto \langle g(x), v\rangle_{\mathcal{Y}}$ is bounded.

The Riesz representation theorem guarantees for each $x \in X$ and $v \in \mathcal{Y}$ the existence of a function $K_{x,v} \in H$ such that $\langle g, K_{x,v}\rangle_H = \langle g(x), v\rangle_{\mathcal{Y}}$ for all $g \in H$. It is readily apparent that the map $K_x : \mathcal{Y} \to H$, that maps v to $K_{x,v}$, is linear, and as such, is expressed as $K_x v := K_{x,v}$.

3. PROBLEM STATEMENT

Given a collection of trajectories, $\{\gamma_i : [0, T] \to \mathbb{R}^n\}_{i=1}^M$, corresponding to a second order dynamical system $\ddot{\gamma} = f(\gamma)$, where f is unknown, we want to determine a model for a trajectory starting at $x(0) = x_0^0$ and $\dot{x}(0) = x_0^1$. In the following, the model is constructed from a finite rank representation of a second order Liouville operator, B_f, obtained via adjoint relations between the B_f, and a collection of occupation kernels.

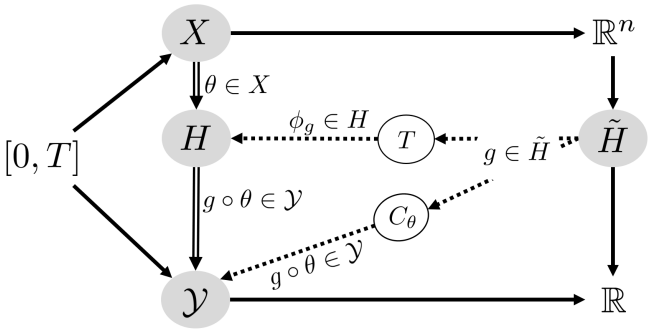


Fig. 1. A visualization of relationships between vector spaces and operators defined in Theorem 1.

4. SIGNAL VALUED RKHSS

In the following, three different RKHSs are under consideration (see Figure 1). The range space, $\mathcal{Y}$, is selected to be a scalar valued RKHS over $[0, T]$, with kernel function $\mathscr{K}$. To construct a vvRKHS of functions that map signals from $C^d([0, T], \mathbb{R}^n)$ (or a suitable substitute) to $\mathcal{Y}$, we define an auxiliary scalar valued RKHS, $\tilde{H}$, consisting of twice continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$. For each $g \in \tilde{H}$, a map from $C^d([0, T], \mathbb{R}^n)$ to $\mathcal{Y}$ is obtained as $\phi_g[\theta](t) := g(\theta(t))$ for all $\theta \in C^d([0, T], \mathbb{R}^n)$ and $t \in [0, T]$. Theorem 1 shows that the space of all such maps is a vvRKHS.

Theorem 1. Let $X = C^d([0, T], \mathbb{R}^n)$ for some $d \in \mathbb{N}$, and let $\tilde{H}$ be a scalar valued RKHS over $\mathbb{R}^n$. Moreover, suppose there exists a RKHS $\mathcal{Y}$ over $[0, T]$ where the composition operator $C_\theta : \tilde{H} \to \mathcal{Y}$ is a bounded operator for all symbols $\theta \in X$. Define the vector space $H(X) := \{\phi_g : g \in \tilde{H}\}$ of mappings $\phi_g : X \to \mathcal{Y}$ given by $\phi_g[\theta] := g(\theta(\cdot))$, together with the inner product induced by $\tilde{H}$, $\langle \phi_{g_1}, \phi_{g_2}\rangle_H = \langle g_1, g_2\rangle_{\tilde{H}}$. Then $H(X)$ is a vvRKHS.

Definition 2. The vvRKHS, H, given in Theorem 1 will be called the signal valued RKHS from $C^d([0, T], \mathbb{R}^n)$ to $\mathcal{Y}$ derived from $\tilde{H}$, more succinctly a signal valued RKHS, when the other quantities are understood from context.

While a general characterization of pairs of RKHSs $(\tilde{H}, \mathcal{Y})$ that admit bounded composition operators C_θ is difficult, the following example analyzes one such pair.

Example 1. A possible choice for $\tilde{H}$ would be the native space of the Gaussian RBF kernel function, $\tilde{K}(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{\mu}\right)$. Letting $\mathcal{Y}$ be the Sobolev space H^1, it can be seen that $\phi_g[\theta] \in \mathcal{Y}$ for all $\theta \in C^2([0, T], \mathbb{R}^n)$. This follows since g is infinitely differentiable and $\phi_g[\theta](t) = g(\theta(t))$ must then be twice continuously differentiable.

Corollary 1. Example 1 provides a vvRKHS of functions from $C^2([0, T], \mathbb{R}^n)$ to H^1, and it is a signal valued RKHS.

DMD relies on the action of Liouville operators on occupation kernels. As such, it is necessary to define second order occupation kernels in the context of vvRKHSs of the form in Theorem 1. To motivate the definition of higher order occupation kernels, recall Cauchy's formula for iterated integrals given as $h^{(-m)}(T) = \frac{1}{(m-1)!}\int_0^T (T-t)^{m-1} h(t)dt$, where $h^{(-m)}(t) := \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_{m-1}} h(\tau_m) d\tau_m \cdots d\tau_2 d\tau_1$.

For a RKHS of continuous functions, $\mathcal{Y}$, as given in Theorem 1, the mapping $h \mapsto h^{(-m)}(T)$ is a bounded functional. As a result, by the Reisz representation theorem, there exists $1^{(-m)} \in \mathcal{Y}$ such that $\langle h, 1^{(-m)}\rangle_{\mathcal{Y}} = \frac{1}{(m-1)!}\int_0^T (T-t)^{m-1}h(t)dt$. Note that for $g \in \tilde{H}$, $\theta \in C^d([0,T],\mathbb{R}^n)$, and $\psi \in H$ such that $\psi = \mathcal{T}g$, the functional

$$\psi \mapsto \langle g\circ\theta, 1^{(-m)}\rangle_{\mathcal{Y}} = \frac{1}{(m-1)!}\int_0^T (T-t)^{m-1}g(\theta(t))dt$$

is bounded. As such, by the Reisz representation theorem, there exists $\Gamma_\theta^{(m)} \in H$ such that $\langle \psi, \Gamma_\theta^{(m)}\rangle_H = \langle g\circ\theta, 1^{(-m)}\rangle_{\mathcal{Y}}$. We define $\Gamma_\theta^{(m)}$ as the *m-th order occupation kernel corresponding to* $\theta \in C^d([0,T],\mathbb{R}^n)$ *in* H.

Due to the fact that $\langle g\circ\theta, 1^{(-m)}\rangle_{\mathcal{Y}} = \langle \psi[\theta], 1^{(-m)}\rangle_{\mathcal{Y}} = \langle \psi, K_{\theta,1^{(-m)}}\rangle_H$, the m-th order occupation kernel corresponding to θ can be identified with the kernel function $K_{\theta,1^{(-m)}} \in H$ of the vvRKHS.

Thus, in contrast with Rosenfeld et al. (2019b), where occupation kernels are integrals of the kernel function of an RKHS along trajectories, the m-th order occupation kernels defined here are a subset of the set of vector valued kernels in a vvRKHS.

5. HIGHER ORDER LIOUVILLE OPERATORS AND OCCUPATION KERNELS

The structure of Liouville operators, given formally as $A_f g(x) = \nabla g(x) f(x)$, derive their form from the orbital derivative. In particular, suppose that $\gamma : [0,T] \to \mathbb{R}^n$ satisfies $\dot\gamma = f(\gamma)$, then $A_f g(\gamma(t)) = \nabla g(\gamma(t))\dot\gamma(t) = \frac{d}{dt}g(\gamma(t))$. Consequently, higher order Liouville operators may be derived via the same process, where g is composed with γ and higher order derivatives with respect to time are taken. To wit, letting $\mathcal{H}[g]$ denote the Hessian of $g : \mathbb{R}^n \to \mathbb{R}$,

$$\frac{d^2}{dt^2} g\circ\gamma(t) = \dot\gamma(t)^T \mathcal{H}[g](\gamma(t))\dot\gamma(t) + \nabla g(\gamma(t))\ddot\gamma(t).$$

Fixing H as a signal valued RKHS as in Theorem 1, let $f : \mathbb{R}^n \to \mathbb{R}^n$ be the symbol for a second order Liouville operator, $B_f : \mathcal{D}(B_f) \to H$, defined as

$$B_f\psi[\theta](t) := \nabla\psi[\theta](t) f(\theta(t)) + \left(\dot\theta(0) + \int_0^t f(\theta(\tau))d\tau\right)^T \mathcal{H}\psi[\theta](t)\left(\dot\theta(0) + \int_0^t f(\theta(\tau))d\tau\right), \quad (1)$$

where $\mathcal{D}(B_f)$ is precisely the collection of ψ for which $B_f\psi \in H$. Note that since $\psi = \phi_g$ for some $g \in \tilde{H}$, $\frac{\partial}{\partial x_i}\psi[\theta](t)$ is defined as $\frac{\partial}{\partial x_i}g(\theta(t))$ for $i = 1,\dots,n$, which facilitates the definitions of the gradient and Hessian of ψ. Hence, when $\ddot\gamma = f(\gamma)$, $B_f\phi_g[\gamma] = \frac{d^2}{dt^2}g\circ\gamma(t)$. Owing to the integral appearing in (1), the operator B_f needs to be posed over a Hilbert space consisting of functions of trajectories. Additionally, in contrast to the first order Liouville operator, B_f is linear in ψ but not in the symbol, f.

The operator B_f is connected to second order occupation kernels in the following manner. If $\ddot\gamma = f(\gamma)$, then $\langle B_f\psi, \Gamma_\gamma^{(2)}\rangle_H = \int_0^T (T-t)B_f\psi[\gamma](t)dt = \int_0^T (T-t)\ddot\psi[\gamma](t)dt = \psi[\gamma](T) - \psi[\gamma](0) - T\nabla\psi[\gamma](0)\dot\gamma(0) = \langle \psi, K_{\gamma,\mathscr{K}_T} - K_{\gamma,\mathscr{K}_0} - TK_{\gamma,\mathscr{K}_0'}\rangle_H$, where $\mathscr{K}_0' := s \mapsto {}^{d(\mathscr{K}(s,t))}/_{dt}|_{t=0} \in \mathcal{Y}$.

Hence, the functional $\psi \mapsto \langle B_f\psi, \Gamma_\gamma^{(2)}\rangle_H$ is bounded, and the following proposition is established.

Proposition 1. Let f be the symbol for a densely defined [1] second order Liouville operator, B_f, over a signal valued RKHS and $\gamma \in C^2([0,T],\mathbb{R}^n)$ be such that $\ddot\gamma = f(\gamma)$. Then, $\Gamma_\gamma^{(2)} \in \mathcal{D}(B_f^*)$, and $B_f^*\Gamma_\gamma^{(2)} = K_{\gamma,\mathscr{K}_T} - K_{\gamma,\mathscr{K}_0} - TK_{\gamma,\mathscr{K}_0'}$.

6. DYNAMIC MODE DECOMPOSITIONS FOR SECOND ORDER DYNAMICAL SYSTEMS

The objective of this section is to give a data driven model for a state governed by an unknown second order dynamical system. The development follows that of occupation kernel DMD detailed in Rosenfeld et al. (2020).

The approach is to determine a finite rank representation of B_f over H and to perform an eigendecomposition on this representation to obtain eigenfunctions and eigenvectors for the representation. Following this, the full state observable is decomposed with respect to the eigenfunctions, which ultimately allows for a model to be extracted for the dynamical system.

Suppose that $\varphi \in \mathcal{D}(B_f)$ is an eigenfunction for B_f with eigenvalue λ. Then for $\ddot\gamma = f(\gamma)$, $\ddot\varphi[\gamma](t) = B_f\varphi[\gamma](t) = \lambda\varphi[\gamma](t)$. Hence,

$$\varphi[\gamma](t) = \frac{1}{2}\left(\varphi[\gamma](0) + \frac{\nabla\varphi[\gamma](0)\dot\gamma(0)}{\sqrt{\lambda}}\right)e^{\sqrt{\lambda}t} + \frac{1}{2}\left(\varphi[\gamma](0) - \frac{\nabla\varphi[\gamma](0)\dot\gamma(0)}{\sqrt{\lambda}}\right)e^{-\sqrt{\lambda}t}.$$

The full state observable for the signal valued case is then given as $\psi_{id}[\theta] = \theta$. The objective is to decompose each dimension of the full state observable with respect to an eigenbasis, $\{\varphi_i\}_{i=1}^\infty$ with eigenvalues $\{\lambda_i\}_{i=1}^\infty$, of B_f, provided that one exists, so as to express $\psi_{id}[\gamma](t)$ as

$$\psi_{id}[\gamma](t) = \gamma(t) = \quad (2)$$

$$\lim_{M\to\infty}\sum_{m=1}^M \xi_{m,M}\left(\frac{1}{2}\left(\varphi_m[\gamma](0) + \frac{\nabla\varphi_m[\gamma](0)\dot\gamma(0)}{\sqrt{\lambda_m}}\right)e^{\sqrt{\lambda_m}t} + \frac{1}{2}\left(\varphi_m[\gamma](0) - \frac{\nabla\varphi_m[\gamma](0)\dot\gamma(0)}{\sqrt{\lambda_m}}\right)e^{-\sqrt{\lambda_m}t}\right). \quad (3)$$

Since the eigenfunctions may not be pairwise orthogonal, addition of each new eigenfunction to the linear combination in (3) may affect the coefficients corresponding to all other eigenfunctions. This dependence of the coefficients on the collection of basis functions is expressed through the second subscript of M. In the following, finite-rank representations of the coefficients $\xi_{m,M}$ are referred to as the second order Liouville modes for the dynamical system.

Since B_f is not known when f is unknown. A finite rank proxy of B_f needs to be constructed from the observed trajectories. In the place of the eigenfunctions of B_f,

[1] An operator $B_f : \mathcal{D}(B_f) \to H$ is called densely defined if $\mathcal{D}(B_f)$ is a dense subset of H.

the eigenfunctions of a finite rank representation will be leveraged to determine an estimate for (3). Let $\{\gamma_i\}_{i=1}^M \subset C^2([0,T],\mathbb{R}^n)$ be a collection of observed trajectories for the second order dynamical system, and let $\alpha = \{\Gamma^{(2)}_{\gamma_i}\}_{i=1}^M$ be the corresponding collection of second order occupation kernels in H. Let P_α be the projection onto $\operatorname{span}\alpha$.

A finite rank representation of B_f restricted to $\operatorname{span}\alpha$, i.e., the matrix $[P_\alpha B_f]_\alpha^\alpha$ that maps the coefficients $\{a_i\}_{i=1}^M$ to the coefficients $\{b_i\}_{i=1}^M$, is given as

$$[P_\alpha B_f]_\alpha^\alpha = \begin{pmatrix} \langle \Gamma^{(2)}_{\gamma_1}, \Gamma^{(2)}_{\gamma_1}\rangle_H & \cdots & \langle \Gamma^{(2)}_{\gamma_1}, \Gamma^{(2)}_{\gamma_M}\rangle_H \\ \vdots & \ddots & \vdots \\ \langle \Gamma^{(2)}_{\gamma_M}, \Gamma^{(2)}_{\gamma_1}\rangle_H & \cdots & \langle \Gamma^{(2)}_{\gamma_M}, \Gamma^{(2)}_{\gamma_M}\rangle_H \end{pmatrix}^{-1} \times \begin{pmatrix} \langle \Gamma^{(2)}_{\gamma_1}, B_f^*\Gamma^{(2)}_{\gamma_1}\rangle_H & \cdots & \langle \Gamma^{(2)}_{\gamma_M}, B_f^*\Gamma^{(2)}_{\gamma_1}\rangle_H \\ \vdots & \ddots & \vdots \\ \langle \Gamma^{(2)}_{\gamma_1}, B_f^*\Gamma^{(2)}_{\gamma_M}\rangle_H & \cdots & \langle \Gamma^{(2)}_{\gamma_M}, B_f^*\Gamma^{(2)}_{\gamma_M}\rangle_H \end{pmatrix}, \tag{4}$$

where B_f was moved to the right of the inner products through the adjoint relation, and the result of the mapping $B_f^*\Gamma^{(2)}_\gamma$ was given in Proposition 1. Letting G denote the Gram matrix $(\langle \Gamma^{(2)}_{\gamma_i}, \Gamma^{(2)}_{\gamma_j}\rangle_H)_{i,j=1}^M$, a normalized "eigenfunction" can be extracted from an eigenvector, ν_j, of $[P_\alpha B_f]_\alpha^\alpha$ with eigenvalue λ_j as

$$\hat{\varphi}_j = \frac{1}{\sqrt{\nu_j^T G \nu_j}} \sum_{i=1}^M (\nu_j)_i \Gamma^{(2)}_{\gamma_i}, \tag{5}$$

which can be leveraged as a proxy for a proper eigenfunction of B_f, in keeping with the implementation of DMD for Koopman and Liouville operators. In (5) and in the following development, $(x)_i$ denotes the projection onto the i-th coordinate of $x \in \mathbb{R}^n$.

The second order Liouville modes can then be constructed by examining the inner products $\left\langle (\psi_{id})_i, \Gamma^{(2)}_{\gamma_j}\right\rangle_H$, where $(\psi_{id})_i$ is the i-th component of the full state observable, i.e., $(\psi_{id})_i[\theta](t) := (\theta(t))_i$. The second order Liouville modes $\{(\xi_m)_i\}_{m=1}^M$ are defined as the coefficients in the projection of $(\psi_{id})_i$ onto the span of the normalized eigenfunctions in (5) that is,

$$\begin{pmatrix} \left\langle (\psi_{id})_1, \Gamma^{(2)}_{\gamma_j}\right\rangle_H \\ \vdots \\ \left\langle (\psi_{id})_n, \Gamma^{(2)}_{\gamma_j}\right\rangle_H \end{pmatrix} \approx \begin{pmatrix} \left\langle \sum_{m=1}^M (\xi_m)_1 \hat{\varphi}_m, \Gamma^{(2)}_{\gamma_j}\right\rangle_H \\ \vdots \\ \left\langle \sum_{m=1}^M (\xi_m)_n \hat{\varphi}_m, \Gamma^{(2)}_{\gamma_j}\right\rangle_H \end{pmatrix}$$
$$= \sum_{m=1}^M \xi_m \sum_{k=1}^M \frac{(\nu_m)_k}{\sqrt{\nu_m^T G \nu_m}} \left\langle \Gamma^{(2)}_{\gamma_k}, \Gamma^{(2)}_{\gamma_j}\right\rangle_H = \sum_{m=1}^M \frac{\xi_m \nu_m^T G^j}{\sqrt{\nu_m^T G \nu_m}},$$

where G^j denotes the j-th column of the Gram matrix. The matrix $\xi := (\xi_1 \cdots \xi_M)$ of second order Liouville modes is then given by

$$\xi = \begin{pmatrix} \left\langle (\psi_{id})_1, \Gamma^{(2)}_{\gamma_1}\right\rangle_H & \cdots & \left\langle (\psi_{id})_1, \Gamma^{(2)}_{\gamma_M}\right\rangle_H \\ \vdots & \ddots & \vdots \\ \left\langle (\psi_{id})_n, \Gamma^{(2)}_{\gamma_1}\right\rangle_H & \cdots & \left\langle (\psi_{id})_n, \Gamma^{(2)}_{\gamma_M}\right\rangle_H \end{pmatrix} \tag{6}$$
$$\times \left(\begin{pmatrix} \frac{\nu_1^T}{\sqrt{\nu_1^T G \nu_1}} \\ \vdots \\ \frac{\nu_M^T}{\sqrt{\nu_M^T G \nu_M}} \end{pmatrix} G \right)^{-1}. \tag{7}$$

REFERENCES

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937.

Budišić, M., Mohr, R., and Mezić, I. (2012). Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4), 047510.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá., V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Anal. Appl.*, 08(01), 19–61.

Kutz, J.N., Brunton, S.L., Brunton, B.W., and Proctor, J.L. (2016). *Dynanic mode decomposition - data-driven modeling of complex systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Lasserre, J.B., Henrion, D., Prieur, C., and Trélat, E. (2008). Nonlinear optimal control via occupation measures and lmi-relaxations. *SIAM Journal on Control and Optimization*, 47(4), 1643–1666.

Li, X. and Rosenfeld, J.A. (to appear). Fractional order system identification with occupation kernel regression. *IEEE Control Systems Letters*. URL `https://ieeexplore.ieee.org/document/9305713`.

Mauroy, A. and Goncalves, J. (2020). Koopman-based lifting techniques for nonlinear systems identification. *IEEE Transactions on Automatic Control*, 65(6), 2550–2565. doi: 10.1109/TAC.2019.2941433.

Mauroy, A. and Mezić, I. (2016). Global stability analysis using the eigenfunctions of the Koopman operator. *IEEE Transactions on Automatic Control*, 61(11), 3356–3369.

Pedrick, G. (1957). *Theory of reproducing kernels for Hilbert spaces of vector valued functions*. Ph.D. thesis, University of Kansas.

Proctor, J.L., Brunton, S.L., and Kutz, J.N. (2016). Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1), 142–161.

Rosenfeld, J., Kamalapurkar, R., Gruss, L.F., and Taylor, J. (2020). Dynamic mode decomposition for continuous time systems with the Liouville operator. arXiv:1910.03977.

Rosenfeld, J.A. and Kamalapurkar, R. (2021, to appear, see arXiv:2101.02620). Dynamic mode decomposition with control Liouville operators. In *Proceedings of the International Symposium on Mathematical Theory of Networks and Systems*.

Rosenfeld, J.A., Kamalapurkar, R., Russo, B., and Johnson, T.T. (2019a). Occupation kernels and densely defined Liouville operators for system identification. In *Proceedings of the IEEE Conference on Decision and Control*, 6455–6460. doi: 10.1109/CDC40024.2019.9029337.

Rosenfeld, J.A., Russo, B., Kamalapurkar, R., and Johnson, T.T. (2019b). The occupation kernel method for nonlinear system identification. arXiv:1909.11792.

Russo, B.P., Kamalapurkar, R., Chang, D., and Rosenfeld, J.A. (2022). Motion tomography via occupation kernels. *Journal of Computational Dynamics*, 9(1), 27–45.

Williams, M.O., Rowley, C.W., and Kevrekidis, I.G. (2015). A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2(2), 247–265.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Accelerating wound healing with feedback control: a data-driven approach ⋆

Marcella M. Gomez *

* *Applied Mathematcs, University of California at Santa Cruz, Santa Cruz, CA 95064 USA (e-mail: mgomez26@ucsc.edu).*

Abstract: Controlling biological systems presents challenges not typically dealt with in traditional control theoretic approaches but also gives way to leniences not traditionally tolerated. Here, we present a holistic view to this new research area and current developments integrating various data-driven approaches for modeling and control.

Keywords: Systems biology, control systems, machine learning, modeling, wound healing

1. INTRODUCTION

Feedback control can help to advance methods in precision medicine (Selberg et al. (2020a)). Feedback control is essential to the regulation of natural biological processes and has been considered as an approach to artificially guide or enhance existing biological systems (e.g. artificial pancreas (El-Khatib et al. (2010); Quiroz (2019)) and neuro-stimulation (Santaniello et al. (2010))).

Realizing feedback control in wound healing requires a way to direct cellular response without genetic engineering. We propose to achieve this through precise control over external signaling cues using bioelectronic devices. To control biological systems, differential voltages are applied to the bioelectronic device in order to drive the delivery or removal of biochemical or biophysical signals to the extracellular environment (Proctor et al. (2019); Malliaras and Abidian (2015); Noy (2015)). These signaling molecules, in turn, drive cellular response.

Bioelectronic devices provide an interface between signal processing and biological tissue that allow one to program custom feedback control strategies with sufficient resolution for enhanced performance. Thus, bioelectronic devices are a promising technology for precision medicine (Löffler et al. (2017); Wu et al. (2017); Birmingham et al. (2014); Selberg et al. (2020a); Jia and Rolandi (2020)). In particular, bioelectronic devices have been at the center of smart bandages (Mostafalu et al. (2018); Farooqui and Shamim (2016)). Many of these bandages have advanced features on board such as sensors to assess the state of wounds in real-time (McLister et al. (2016); Sharp et al. (2010)) and controlled release of therapeutics in a variety of pathological conditions (Williamson et al. (2015); Proctor et al. (2019); Jonsson et al. (2015)). Here, we propose to advance the capabilities of smart bandages with feedback control.

Methods in control theory typically assume that a predictive model is available. Thus, the models can be used to determine the most suitable type of controller and tune parameters for the controller. The challenges that we face in feedback control ensue from our efforts to control a complex systems for which we do not have a predictive model and there are limited observable states. Additionally, we are trying to achieve a goal at the tissue level response by controlling biological processes at the single cell level. Thus, we need a way of mapping single cell response on short timescales to tissue level response over the course of wound healing.

In summary, we propose that an effective approach to controlling complex biological processes interfaced with a bioelectronic device is through a heirarchical control architecture. The heirarchical feedback control architecture allows one to design the components of the controller independently and their integration provides control objectives at the single-cell level to achieve a desired tissue level response. That is, our controller design at the single-cell level does not have any dependence on the wound healing model. The wound healing model is instead used to inform the desired wound environment to be achieved by the bioelectronic device at the single-cell level. We note that the time scale of the dynamical response of the bioelectronic device to changes in voltages is orders of magnitude faster than that of biological processes in wound healing. We propose the following three layers in this heirarchical structure: the Decision Maker, Planner, and Low-level Controller. We also argue that a data-driven approach allows us to be successful without a complete and predictive mechanistic model of wound healing. Below, we describe our approach in more detail.

2. MACHINE LEARNING FOR COMPLEX SYSTEMS

Machine learning (ML)-based techniques are suitable when accurate control is required in the absence of a precise mathematical model (Marquez et al. (2019)). The best-

⋆ Research was sponsored by the Office of Naval Research and the DARPA Biotechnologies Office (DARPA/BTO) and was accomplished under Cooperative Agreement Number DC20AC00003. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research and the DARPA Biotechnologies Office (DARPA/BTO) or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

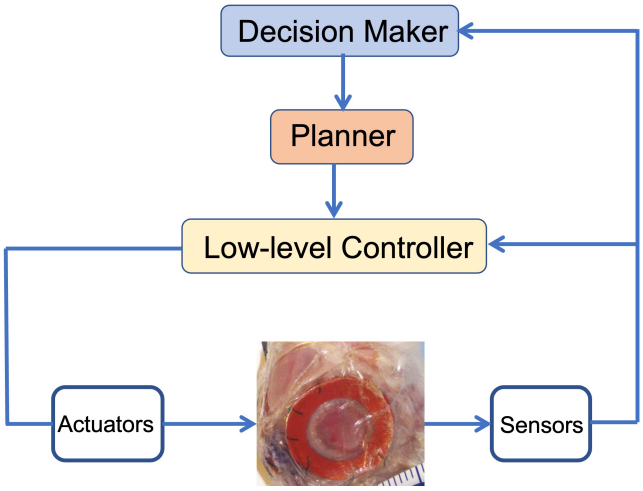


Fig. 1. Heirarchical control scheme.

known ML techniques rely on the availability of large datasets a priori and have not been applied to control bioelectronic devices (Angermueller et al. (2016); Camacho et al. (2018); Maltarollo et al. (2013); Park and Kellis (2015)). We propose that ML-based techniques that are explored as control solutions outside of biology for cases involving complex non-linear systems are also suitable for closing the loop for bioelectronic systems containing biosensors, biology, and bioelectronic actuators. To this end, tools from control systems theory leveraging ML can be used to learn from new observations for effective real-time operation without data a priori (Jafari and Gomez (2019); Jafari et al. (2020); Kumpati et al. (1990); Lavretsky and Wise (2013)). ML-based techniques can be implemented directly or indirectly to solve complex control problems (Hagan and Demuth (1999); Spooner et al. (2004)). Additionally, we leverage ML to build data-driven models for long term trajectory planning.

3. SYSTEM OF SYSTEMS APPROACH

3.1 Low-level Controller

The low-level controller directly drives cellular response. The architecture of this inner loop resembles a standard feedback control loop composed of the system, sensors, actuators, and controller. A reference trajectory must be provided to the low-level controller. The controller is then tasked with achieving the desired response based on the sensor readout. In this work, we have explored various types of controllers including a NN-based controller with no information a priori and a sliding mode controller that can deal with saturating controller outputs imposed by the limitations of the bioelectronic device.

We have already demonstrated closed-loop control of membrane voltage (Vmem) in human pluripotent stem cells using bioelectronic stimulation (Selberg et al. (2020b)). This type of closed loop control relies on fluorescent readout of cell state. This state is then fed into a NN-based controller that suggests to the bioelectronic device the necessary intervention to achieve the desired state, in this specific example in the form of H+ ions or pH (Jia et al. (2020b)). Unique to this approach is the ability to obtain spatial resolution close to the single cell level (100um) and temporal resolution than spans many timescales from short cell-events (ms) to cell development (hrs and days). This approach goes well beyond Vmem and H+ ions and can be used to control various types of cell behavior (Selberg et al. (2020a, 2018)), including fate, with many types of bioelectronic signals that include ions (Jia et al. (2020a)), small molecules (Poxson et al. (2019)), and electric field.

3.2 Planner

The Planner takes in high level instructions from the Decision Maker and generates reference signals and/or control objectives for the Low-level Controller. In recent work, we have established a qualitative reduced order model of wound healing to understand how timing of macrophage polarization affects overall wound healing time (Zlobina et al. (2021)). Macrophages play an essential role in wound healing. Additionally, there are two primary subtypes termed M1 and M2 macrophages. M1 macrophages are pro-inflammatory and M2 are anti-inflammatory and promote the onset of the proliferative stage of wound healing. The transition and timing of M1 to M2 macrophages is critical to ensuring healthy wound healing. We aim to manipulate the timing and speed of this process in order to reduce inflammation time and accelerate transition into proliferation, thereby, reducing wound healing time.

3.3 Decision Maker

The Decision maker consists of a high-level model of the wound healing process. In particular, this component of the controller is tasked with identifying and monitoring progression through the wound healing stages in order to accurately time interventions designed by the Planner. We present preliminary ML-based models that predict wound healing stage based on gene expression profiles and time series RGB images of the wound.

4. CONCLUSION

In conclusion, we propose a framework for controlling biological systems motivated by our goal to accelerate wound healing. This framework has the potential to be generalized and extendended to other applications in biology such as precision medicine, agriculture, and environmental health, where systems are high dimensional and dynamics are complex. Still some open questions remain such as controllability of the system and safety guarantees.

ACKNOWLEDGEMENTS

The ongoing effort is part of a large collaborative project involving PIs, research scientists, postdocs, and students across various univerisities, departments, and disciplines. The work described in this extended abstract is led by the author but carried out over the course of the project by various lab members. The author would like to give thanks to research scientist Ksenia Zlobina, to former postdocs Mohammad Jafari and Bashir Hosseini Jafari, graduate students Giovanny Marquez, Sam Teymoori, Hector Carrion, and Manasa Kesapragada. Additionally, the author would like to thank former masters students who have contributed to the effort including Eliana Phillips, Brett Sargent, Jiahao Xue, Krishnakant Dasika, and undergraduate reearcher Han Chen.

REFERENCES

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, 12(7), 878.

Birmingham, K., Gradinaru, V., Anikeeva, P., Grill, W.M., Pikov, V., McLaughlin, B., Pasricha, P., Weber, D., Ludwig, K., and Famm, K. (2014). Bioelectronic medicines: a research roadmap. *Nature Reviews Drug Discovery*, 13(6), 399–400.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-generation machine learning for biological networks. *Cell*, 173(7), 1581–1592.

El-Khatib, F.H., Russell, S.J., Nathan, D.M., Sutherlin, R.G., and Damiano, E.R. (2010). A bihormonal closed-loop artificial pancreas for type 1 diabetes. *Science translational medicine*, 2(27), 27ra27–27ra27.

Farooqui, M.F. and Shamim, A. (2016). Low cost inkjet printed smart bandage for wireless monitoring of chronic wounds. *Sci Rep*, 6, 28949. doi:10.1038/srep28949. URL `https://www.ncbi.nlm.nih.gov/pubmed/27353200`.

Hagan, M.T. and Demuth, H.B. (1999). Neural networks for control. In *Proceedings of the 1999 American control conference (cat. No. 99CH36251)*, volume 3, 1642–1656. IEEE.

Jafari, M. and Gomez, M. (2019). Online machine learning based controller for coupled tanks systems. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 163–169. IEEE.

Jafari, M., Marquez, G., Selberg, J., Jia, M., Dechiraju, H., Pansodtee, P., Teodorescu, M., Rolandi, M., and Gomez, M. (2020). Feedback control of bioelectronic devices using machine learning. *IEEE Control Systems Letters*, 5(4), 1133–1138.

Jia, M., Dechiruji, H., Selberg, J., Pansodtee, P., Mathews, J., Wu, C., Levin, M., Teodorescu, M., and Rolandi, M. (2020a). Bioelectronic control of chloride ions and concentration with ag/agcl contacts. *APL Materials*, 8(9), 091106.

Jia, M., Ray, S., Breault, R., and Rolandi, M. (2020b). Control of ph in bioelectronics and applications. *APL Materials*, 8(12), 120704.

Jia, M. and Rolandi, M. (2020). Soft and ion-conducting materials in bioelectronics: From conducting polymers to hydrogels. *Advanced healthcare materials*, 9(5), 1901372.

Jonsson, A., Song, Z., Nilsson, D., Meyerson, B.A., Simon, D.T., Linderoth, B., and Berggren, M. (2015). Therapy using implanted organic bioelectronics. *Sci Adv*, 1(4), e1500039. doi:10.1126/sciadv.1500039. URL `https://www.ncbi.nlm.nih.gov/pubmed/26601181`.

Kumpati, S.N., Kannan, P., et al. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on neural networks*, 1(1), 4–27.

Lavretsky, E. and Wise, K.A. (2013). Robust adaptive control. In *Robust and adaptive control*, 317–353. Springer.

Löffler, S., Melican, K., Nilsson, K., and Richter-Dahlfors, A. (2017). Organic bioelectronics in medicine. *Journal of internal medicine*, 282(1), 24–36.

Malliaras, G. and Abidian, M.R. (2015). Organic bioelectronic materials and devices. *Advanced materials (Deerfield Beach, Fla.)*, 27(46), 7492.

Maltarollo, V.G., Honório, K.M., and da Silva, A.B.F. (2013). Applications of artificial neural networks in chemical problems. *Artificial neural networks-architectures and applications*, 203–223.

Marquez, G., Johnson, B., Jafari, M., and Gomez, M. (2019). Online machine learning based predictor for biological systems. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 120–125. IEEE.

McLister, A., McHugh, J., Cundell, J., and Davis, J. (2016). New developments in smart bandage technologies for wound diagnostics. *Advanced Materials*, 28(27), 5732–5737.

Mostafalu, P., Tamayol, A., Rahimi, R., Ochoa, M., Khalilpour, A., Kiaee, G., Yazdi, I.K., Bagherifard, S., Dokmeci, M.R., Ziaie, B., Sonkusale, S.R., and Khademhosseini, A. (2018). Smart bandage for monitoring and treatment of chronic wounds. *Small*, e1703509. doi:10.1002/smll.201703509. URL `https://www.ncbi.nlm.nih.gov/pubmed/29978547`.

Noy, A. (2015). Mimicking biology with nanomaterials: carbon nanotube porins in lipid membranes. *Biophysical Journal*, 108(2), 443a.

Park, Y. and Kellis, M. (2015). Deep learning for regulatory genomics. *Nature biotechnology*, 33(8), 825–826.

Poxson, D.J., Gabrielsson, E.O., Bonisoli, A., Linderhed, U., Abrahamsson, T., Matthiesen, I., Tybrandt, K., Berggren, M., and Simon, D.T. (2019). Capillary-fiber based electrophoretic delivery device. *ACS applied materials & interfaces*, 11(15), 14200–14207.

Proctor, C.M., Chan, C.Y., Porcarelli, L., Udabe, E., Sanchez-Sanchez, A., Del Agua, I., Mecerreyes, D., and Malliaras, G.G. (2019). Ionic hydrogel for accelerated dopamine delivery via retrodialysis. *Chem Mater*, 31(17), 7080–7084. doi:10.1021/acs.chemmater.9b02135. URL `https://www.ncbi.nlm.nih.gov/pubmed/32063677`.

Quiroz, G. (2019). The evolution of control algorithms in artificial pancreas: A historical perspective. *Annual Reviews in Control*, 48, 222–232.

Santaniello, S., Fiengo, G., Glielmo, L., and Grill, W.M. (2010). Closed-loop control of deep brain stimulation: a simulation study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(1), 15–24.

Selberg, J., Jafari, M., Bradley, C., Gomez, M., and Rolandi, M. (2020a). Expanding biological control to bioelectronics with machine learning. *APL Materials*, 8(12), 120904.

Selberg, J., Gomez, M., and Rolandi, M. (2018). The potential for convergence between synthetic biology and bioelectronics. *Cell systems*, 7(3), 231–244.

Selberg, J., Jafari, M., Mathews, J., Jia, M., Pansodtee, P., Dechiraju, H., Wu, C., Cordero, S., Flora, A., Yonas, N., et al. (2020b). Machine learning-driven bioelectronics for closed-loop control of cells. *Advanced Intelligent Systems*, 2(12), 2000140.

Sharp, D., Gladstone, P., Smith, R.B., Forsythe, S., and Davis, J. (2010). Approaching intelligent infection diagnostics: Carbon fibre sensor for electrochemical pyocyanin detection. *Bioelectrochemistry*, 77(2), 114–119.

Spooner, J.T., Maggiore, M., Ordonez, R., and Passino, K.M. (2004). *Stable adaptive control and estimation for nonlinear systems: neural and fuzzy approximator techniques*. John Wiley & Sons.

Williamson, A., Rivnay, J., Kergoat, L., Jonsson, A., Inal, S., Uguz, I., Ferro, M., Ivanov, A., Sjostrom,

T.A., Simon, D.T., Berggren, M., Malliaras, G.G., and Bernard, C. (2015). Controlling epileptiform activity with organic electronic ion pumps. *Adv Mater*, 27(20), 3138–44. doi:10.1002/adma.201500482. URL `https://www.ncbi.nlm.nih.gov/pubmed/25866154`.

Wu, H., Gao, W., and Yin, Z. (2017). Materials, devices and systems of soft bioelectronics for precision therapy. *Advanced healthcare materials*, 6(10), 1700017.

Zlobina, K., Xue, J., and Gomez, M. (2021). Effective spatio-temporal regimes for wound treatment by way of macrophage polarization: A mathematical model.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Neural Networks Approximations for State Constrained Deterministic Control ⋆

.

Olivier Bokanowski *, **Averil Prost** **, **Xavier Warin** ***

* *Lab. Jacques-Louis Lions, Université Paris Cité, Paris, France (e-mail: olivier.bokanowski@math.univ-paris-diderot.fr)*
** *Lab. Jacques-Louis Lions, Université Paris Cité, Paris, France (e-mail: averil.prost@insa-rouen.fr)*
*** *EDF R&D, FiME (e-mail: xavier.warin@edf.fr)*

Abstract: We propose new neural networks algorithms for the approximation of deterministic optimal control problems with maximum running cost. This problem is motivated by the approximation of general optimal control problems in the presence of state constraints. This problem is also related to Hamilton-Jacobi-Bellman equations with an obstacle term. Difficulties arise in particular because of the non-smoothness of the value to be approximated, and appropriate solutions are studied to deal with this specific issue. Numerical examples are given on front propagation problems in the presence of an obstacle, for average dimensions $2 \leq d \leq 8$.

Keywords: deterministic optimal control, maximum running cost, high dimension, deep learning, Hamilton-Jacobi-Bellman equations, state-constraints, obstacle problem, level-set

GENERAL PRESENTATION

In this work (see Bokanowski et al. (Preprint) for details) we present new results and algorithms concerning the use of deep neural network (DNN) approximations for a deterministic finite-horizon optimal control problem in presence of state constraints.

We focus on the value corresponding to a maximum running cost problem with obstacle function $g(.)$, for a given $T > 0$, $t \in [0, T]$ and $x \in \mathbb{R}^d$:

$$v(t,x) = \inf_{a \in mes((0,T),A)} \max_{\theta \in (t,T)} g(y^a_{t,x}(\theta)) \bigvee \varphi(y^a_{t,x}(T)) \quad (1)$$

where $y(.) = y^a_{t,x}(.)$ obeys $\dot{y}(s) = f(y(s), a(s))$ a.e. $s \in (t,T)$ with $y(t) = x$, associated to a measurable control function $a : (0,T) \to A$, $A \subset \mathbb{R}^\kappa$ is a compact set, and $g(.), \varphi(.), f(.,.)$ are assumed Lipschitz continuous.

This problem is motivated by the computation of backward reachable set with state constraints (see Bokanowski et al. (2010)). Moreover, this framework can be used in order to compute the value of general deterministic optimal control problems with a running cost, terminal cost, and state constraints, following Altarovici et al. (2013): an auxiliary value problem with one more variable enables to deal with the state constraints and avoids discontinuous value functions.

This setting was generalized in Germain et al. (2021) for the control of state-constrained McKean-Vlasov equations (using DNN approximations). It is also applied for *optimistic control* in Bokanowski et al. (2022).

⋆ This research benefited from the support of the FMJH program PGMO and from the FiME laboratory

The value is also the solution of an Hamilton-Jacobi-Bellman (HJB) equation in presence of an obstacle term

$$\begin{aligned} &\min(v_t + H(x, \nabla_x v),\ v - g(x)) = 0, \quad t \in [0,T],\ x \in \mathbb{R}^d \\ &v(T,x) = \max(\varphi(x), g(x)), \quad x \in \mathbb{R}^d \end{aligned} \quad (2)$$

hence an approximation for (1) is also valid for (2).

Recently, deep neural network (DNN) approximations for control in a probabilistic context have been introduced in Han et al. (2018) (deep BSDE algorithm), and also in Huré et al. (2021) and Bachouch et al. (2022), for the approximation of stochastic control problems on finite horizon.

We are motivated in dealing with state-constrained deterministic control problems of average state dimension (such as e.g. 5 to 10) where a brute force discretization of the HJB equation (or of the dynamic programming principle) is in general intractable or too costly to consider. In this framework, in the absence of diffusion terms, the value function is less regular than the usual setting for DNN approximations schemes. In general the presence of state constraints also prevents the value to be regular (because of the possible non-existence of feasible trajectories). Then the DNN may fail to well approximate the desired value function. From a theoretical point of view, the known convergence results for deep neural network (DNN) approximations use at some the level the presence of diffusion. The approach in Huré et al. (2021) utilizes that the law of the process has a density, which is not the case for deterministic evolution equations. Also, the general convergence result for deep BSDE in Han and Long (2020) needs a diffusion assumption on the SDE model as well as restrictive assumptions on the dynamics.

Here, we investigate, in the deterministic setting and by using a level-set approach, the use of DNN schemes based on the approximation of the dynamic programming principle. We develop new schemes extending some ideas coming from stochastic control of Huré et al. (2021), Bachouch et al. (2022) and Germain et al. (2022). These schemes are based on a time discretization of the PDE and some time local optimizations using classical feedforward networks (in our experiments, global methods such as the deep Galerkin method of Sirignano and Spiliopoulos (2018) may fail to see the obstacle).

We are able to prove the convergence of the algorithm in some L^1 norm. We also illustrate numerically the potential of the algorithm (and variants) on some academic front propagation problems in presence of obstacles in average dimensions (e.g. 2 to 8).

IDEA OF THE RESULT

We will use an equivalent formulation of Bellman's dynamic programming principle using feedback controls $a \in \mathcal{A} := mes(\mathbb{R}^d, A)$, the set of measurable functions from $\mathbb{R}^d$ into A. For a given $N \geq 1$, let $\Delta t = \frac{T}{N}$, $t_k = k\Delta t$, and consider approximate characteristics with time step Δt, denoted $F^a(x)$, for the approximation of $\dot{y}(t) = f(y(t), a(y(t)))$, $y(t_n) = x$, $t \in [t_n, t_{n+1}]$. For instance, the Euler scheme is

$$F^a(x) = x + \Delta t f(x, a(x));$$

the Heun scheme in our setting reads

$$F^a(x) = x + \frac{\Delta t}{2}(f(x, a(x) + f(x_1, a(x))$$

where $x_1 = x + \Delta t f(x, a(x))$ (here the control is frozen at the value $a(x)$ where x is the foot of the characteristic), etc. More complex explicit or implicit RK schemes have also to be considered (see Bokanowski et al. (Preprint)). The problem is then to compute an approximation of

$$V_n(x) := \min_{(a_n,\dots,a_{N-1})\in\mathcal{A}^{N-n}} \Big(\max_{k=n,\dots,N} g(X^a_{k,x}) \Big) \bigvee \varphi(X^a_{N,x})$$

where, for $a = (a_n, \dots, a_{N-1})$, the discrete dynamics $(X^a_{k,x})_{k=n,\dots,N}$ is such that

$$\begin{aligned} &X^a_{n,x} := x \\ &X^a_{k+1} := F^{a_k}(X^a_{k,x}), \quad \forall k = n, \dots, N-1. \end{aligned}$$

One of the considered scheme (the "Lagragian scheme") in a simplified form, is as follows. Let $(\hat{A}^\Theta_n)_{n\in[\![0,N-1]\!]}$ be a given sequence of finite-dimensional spaces (such as feedforward neural networks), for the approximation of feedback controls. We also consider a sequence of random variables $(X_k)_{k=0,\dots,N}$ with associated densities $\rho_k \in L^1(\mathbb{R}^d)$ (assuming for instance $\rho_k > 0$).

Algorithm ("Lagrangian scheme") *Set $\hat{V}_N := g \vee \varphi$. Then, for $n \in N-1, \dots, 0$, compute $\hat{a}_n \in \hat{A}^\Theta_n$ and set $\hat{V}_n$ as follows:*

$$\hat{a}_n \in \operatorname*{argmin}_{a\in\hat{A}^\Theta_n} \mathbb{E}\Big[g(X_n) \bigvee \hat{V}_{n+1}(F^a(X_k)) \Big] \tag{3a}$$

$$\hat{V}_n(x) := g(x) \bigvee \hat{V}_{n+1}\left(F^{\hat{a}_k}(x)\right) \tag{3b}$$

In this algorithm, only the feedback controls $(\hat{a}_k)$ are stored ($\hat{V}_n$ is not stored). In practice, the minimization problem (3a) is dealt with a stochastic gradient method. Each evaluation of the value $\hat{V}_{n+1}(x)$ uses the previous controls $(\hat{a}_{n+1}, \dots, \hat{a}_{N-1})$ to compute the approximated characteristic, in a full Lagrangian philosophy. Then we can show a convergence result in average, of the form

$$\max_{0\leq k\leq N} \mathbb{E}[|\hat{V}_k(X_k) - V_k(X_k)|] \overset{\Theta\to\infty}{\to} 0$$

as the parameters Θ for the neural network approximation space grows to infinity that is, assuming that

$$\max_{0\leq k\leq N-1} \inf_{a_k\in\hat{A}^\Theta_k} \mathbb{E}[|a_k(X_k) - \bar{a}_k(X_k)|] \overset{\Theta\to\infty}{\to} 0,$$

$(\bar{a}_k)_{0\leq k\leq n}$ beeing given optimal feedback controls for V_0. A difficulty in showing the convergence is the lack of regularity of the controls in feedback form, which are in general discontinuous. In order to deal with this issue, we first construct near optimal Lipschitz continuous feedback controls in order to approximate the value in some average norm; then we show that our algorithm can produce approximations of these controls (and associated values) up to arbitrary precision, therefore leading to a global convergence result for the approximated values.

REFERENCES

Altarovici, A., Bokanowski, O., and Zidani, H. (2013). A general Hamilton-Jacobi framework for non-linear state-constrained control problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(337–357).

Bachouch, A., Huré, C., Langrené, N., and Pham, H. (2022). Deep Neural Networks Algorithms for Stochastic Control Problems on Finite Horizon: Numerical Applications. *Methodol. Comput. Appl. Probab.*, 24(1), 143–178.

Bokanowski, O., Forcadel, N., and Zidani, H. (2010). Reachability and minimal times for state constrained nonlinear problems without any controllability assumption. *SIAM J. Control Optim.*, 48(7), 4292–4316. doi:10.1137/090762075. URL `http://dx.doi.org/10.1137/090762075`.

Bokanowski, O., Gammoudi, N., and Zidani, H. (2022). Optimistic Planning Algorithms For State-Constrained Optimal Control Problems. *Computers & Mathematics with Applications*, 109(1), 158–179.

Bokanowski, O., Prost, A., and Warin, X. (Preprint). Neural networks for deterministic HJB equations and application to front propagation with obstacle terms.

Germain, M., Pham, H., and Warin, X. (2021). A level-set approach to the control of state-constrained McKean-Vlasov equations: application to renewable energy storage and portfolio selection. URL `https://hal.archives-ouvertes.fr/hal-03498263`. Preprint.

Germain, M., Pham, H., and Warin, X. (2022). Approximation error analysis of some deep backward schemes for nonlinear pdes. *SIAM Journal on Scientific Computing*, 44(1), A28–A56.

Han, J., Jentzen, A., and E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA*, 115(34), 8505–8510.

Han, J. and Long, J. (2020). Convergence of the deep BSDE method for coupled FBSDEs. *Probab. Uncertain. Quant. Risk*, 5, Paper No. 5, 33.

Huré, C., Pham, H., Bachouch, A., and Langrené, N. (2021). Deep neural networks algorithms for stochastic control problems on finite horizon: convergence analysis. *SIAM J. Numer. Anal.*, 59(1), 525–557.

Sirignano, J. and Spiliopoulos, K. (2018). Dgm: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375, 1339–1364.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A mathematical model for depression

Björn S. Rüffer *

* *Bauhaus-Universität Weimar, 99421 Weimar, Germany*
(e-mail: bjoern.rueffer@uni-weimar.de)

Abstract: A basic dynamical model for (clinical) depression is proposed that describes the time evolution of two coupled states: a depression symptom and the memory of past symptoms. The model consists of a system of two coupled first order differential equations with unit coefficients that qualitatively captures different courses of illness.

Keywords: Clinical depression, dynamical system, ordinary differential equations, stability analysis.

1. INTRODUCTION

According to the World Health Organization (WHO), depression is a common mental disorder affecting 5% of the adult population and a leading cause of disability worldwide, and can lead to suicide. The WHO further estimates that over 700 000 people die due to suicide every year, with suicide being the fourth leading cause of death in 15–29 year-olds (World Health Organization, 2021).

While the WHO also states that "there is effective treatment for mild, moderate, and severe depression", anecdotal evidence observed by the author has it that in some individuals depression can be very difficult to treat, has temporally fluctuating features, and psychiatric treatment often involves a long trial-and-error process for finding an effective medication regime that strikes an acceptable balance between effectiveness and its often severe side effects.

Clinical depression, also known as major or severe depression, is the stronger form of depression, and its treatment is more involved than simply administering the right drug (after identifying the right one in the first place). Research on the development of medical drugs would naturally involve neuroscience and neurobiology as well as medical studies.

At the same time, there exist a wealth of psychological research and established therapies for treating depression without administering drugs, including behavioral therapy (attributed independently to Wolpe; to Skinner; as well as to Rachman and Eysenck) and Aaron T. Beck's cognitive therapy and cognitive behavioral therapy.

There exist mathematical models in both of these reasonably distinct branches of depression research, e.g., Byrum et al. (1999); Disner et al. (2011). A common feature that can often be seen on both sides consists of causality networks that show how signaling pathways and areas in the brain influence each other, or how schemas, memories, triggers and behaviors are coupled in different psychological theories. An aspect most often overlooked, or at least not studied at the same level of detail, is the time axis, and the dynamical behavior of these graph-based models.

It is important to note that unlike in mathematics or the exact sciences, it is practically impossible to either derive a model from first principles or to identify a model from data, and that a model that is "as simple as possible but not any simpler" is the most desirable model. Another difficulty is that it is often unclear whether the root source of a patient's depression is genetic or a result of external factors, or a combination.

In the present contribution we propose a dynamical model that only implements the most basic coupling topology. Our model is simple in terms of its mathematical description, but not too simple to not exhibit complex enough behaviors consistent with anecdotal observations about the course of a depression.

There are two states in our model. One state represents the severity of depression or another suitable symptom that is a good indicator of depression. This is something that would commonly be measured by a practitioner using a questionnaire, such as the Beck Depression Inventory-Second Edition (BDI-2), a widely used 21-item self-report inventory measuring the severity of depression in adolescents and adults (Beck et al., 1996), or the Beck Hopelessness Scale (BHS), an instrument for assessing cognitive thoughts among suicidal persons (Beck et al., 1988). The second state measures the memory of past depression experiences, following the schema idea in Beck's cognitive theory (Disner et al., 2011). In a nutshell, this memory represents internal belief representations of past negative stimuli or experiences, which has influence on the effect of new stimuli. Unlike the depression state, the memory state evades direct measurement and for a given individual may only be inferred by the course of the illness in relation to external stimuli, which are just as difficult to measure and quantify.

The two states are coupled, and higher values of the memory state effect that external stimuli (which commonly would consist of negative life events for the individual) more severely worsen the depression, thus increase the depression state. Persistently large values of the depression state on the other hand effect a stronger increase of the memory state, while in the absence of depression symptoms, the memory state attenuates over time.

In Section 2 we propose a mathematical model from the informal description given above, along with some design choices for the numerical range of the states and the nature of the coupling. As our modeling aims to be qualitative and not to quantitatively match a given individual, there is no loss in generality to assume that all coupling coefficients have unit value. In Section 3 we examine properties of the model and study several scenarios in some detail. Section 4 concludes this study and provides some ideas for future research.

2. THE MATHEMATICAL MODEL OF DEPRESSION

Without loss of generality, we can normalize both the depression state (according to a scale such as BDI-2 or BHS, which naturally have a bounded range of possible values), which we call s, and the memory state, which we denote by m, to the unit interval $[0, 1]$, with zero denoting the absence of depression or a memory of it, and one the most severe form of depression and maximal memory of it.

We denote the external stimulus or influence to the model by $e \geq 0$, with the convention that zero means absence of the stimulus and larger values mean more severe and pronounced forms of negative life events affecting the individual. This stimulus may change with time.

With this notation the evolution of the symptom is given by

$$\dot{s} = \big(e(1 + s + m) - s(1 - m)\big)(1 - s)s \tag{1}$$

where the second and third factors, $(1-s)$ and s, are simply there to confine the solutions to the interval $[0, 1]$. The growth term $e(1+s+m)$ is always non-negative. It models that the symptom increases if there is an external influence (non-zero e), and that this rate of change is increased by the combined effect of current symptom severity (s) as well as symptom history (m). The decay term $-s(1 - m)$ is always non-positive and models the attenuation of the symptom. The net change rate is the combination of the growth and decay terms.

The evolution of the memory is given by

$$\dot{m} = \big(s - (1 - s)m\big)(1 - m)m. \tag{2}$$

Again the second and third factor ensure that the memory remains confined to the interval $[0, 1]$. The growth term in the change rate is simply s, the symptom severity, while the decay term is $-(1 - s)m$, modeling an an attenuation of the the memory that is stronger if symptoms are low, and less pronounced if symptoms are more developed.

Denoting the tuple $x := (x_1, x_2)^T := (s, m)^T \in [0, 1]^2$, we can write the model in the more compact form

$$\dot{x} = f(x, e) \tag{3}$$

with

$$f(x, e) = \begin{pmatrix} \big(e(1 + x_1 + x_2) - x_1(1 - x_2)\big)(1 - x_1)x_1 \\ \big(x_1 - (1 - x_1)x_2\big)(1 - x_2)x_2 \end{pmatrix}.$$

Expanding f, it is apparent that the model is nonlinear (polynomial in fact), which allows it to accommodate a rich set of dynamics as we shall see next.

3. PROPERTIES OF THE MODEL OF DEPRESSION

3.1 Equilibria

Using $x = (x_1, x_2)^T$ and $(s, m)^T$ interchangeably and treating $e \geq 0$ as a fixed parameter, we find the following list of equilibrium points for system (3):

$$\begin{aligned}
\bar{x}_1 &= (0, 0)^T \\
\bar{x}_2 &= (-\tfrac{e}{e-1}, 0)^T \\
\bar{x}_3 &= (1, 0)^T \\
\bar{x}_4 &= (\tfrac{e+\sqrt{5\,e^2-10\,e+1}-1}{2\,(e-2)}, -\tfrac{3\,e+\sqrt{5\,e^2-10\,e+1}-1}{2\,(e+1)})^T \\
\bar{x}_5 &= (\tfrac{e-\sqrt{5\,e^2-10\,e+1}-1}{2\,(e-2)}, -\tfrac{3\,e-\sqrt{5\,e^2-10\,e+1}-1}{2\,(e+1)})^T \\
\bar{x}_6 &= (0, 1)^T \\
\bar{x}_7 &= (-2, 1)^T \\
\bar{x}_8 &= (1, 1)^T
\end{aligned}$$

Of these, equilibrium point $\bar{x}_7$ is clearly outside our state space $[0, 1]^2$ and can be ignored in subsequent investigations.

Equilibria $\bar{x}_2$, $\bar{x}_4$, and $\bar{x}_5$ are dependent on the input e, which may vary with time.

Equilibria $\bar{x}_1$, $\bar{x}_3$, $\bar{x}_6$, and $\bar{x}_8$ are by construction of the model, while the point $\bar{x}_2$ is well-defined and in the domain $[0, 1]^2$ only if $e < 1$.

Provided that $e \leq 1 - \frac{2}{5}\sqrt{5} \approx 0.1$, the equilibria at $\bar{x}_4$ and $\bar{x}_5$ are well-defined and located in $[0, 1]^2$.

Already at this point it is clear that the model will exhibit changes in dynamics when the value of e varies across the value of $1 - \frac{2}{5}\sqrt{5} \approx 0.1$ or about 1.

It is an artifact of the design choices in this model that the points $\bar{x}_1$, $\bar{x}_3$, $\bar{x}_6$, $\bar{x}_8$ are equilibria irrespective of the value of e, meaning that an individual who has, say, zero symptoms and memory, will always remain this way, unfazed by any external stimuli. For practical purposes, however, it is a more realistic scenario to consider whether the states asymptotically converge to zero, or remain close to zero, which is a question of (asymptotic) stability.

Put differently, no real world and alive individual can ever be found in states $\bar{x}_1$ or $\bar{x}_8$, as everyone will have had some negative experiences in their lives.

3.2 Linearization and local stability analysis

Linearizing f at x we obtain the Jacobian $J(x, e) := \frac{\partial f}{\partial x}(x, e)$. At every point $x \in [0, 1]^2$ the matrix $J(x, e)$ is Metzler, i.e., it has non-negative off-diagonal entries. This means that the system $\dot{x} = f(x, e)$ is monotone (in x), which further implies that it is positive (i.e., positive initial conditions result in states that remain positive) and solutions for initial condition in $[0, 1]^2$ and arbitrary non-negative and locally integrable input $e = e(t)$ exist for all times and are confined to $[0, 1]^2$, see, e.g., Smith (1995) for an introduction to monotone systems. We'll discuss monotonicity in more detail in Section 3.4.

Evaluating the Jacobian of f at the equilibrium points in the four corners of the state space, we find

- $J(\bar{x}_1) = \begin{pmatrix} e & 0 \\ 0 & 0 \end{pmatrix}$,
- $J(\bar{x}_3) = \begin{pmatrix} -2\,e+1 & 0 \\ 0 & 1 \end{pmatrix}$,
- $J(\bar{x}_6) = \begin{pmatrix} 2\,e & 0 \\ 0 & 1 \end{pmatrix}$, and
- $J(\bar{x}_8) = \begin{pmatrix} -3\,e & 0 \\ 0 & -1 \end{pmatrix}$.

We can immediately read off the eigenvalues from the diagonals and conclude that

- equilibrium $\bar{x}_1$ is unstable for $e > 0$; no conclusion can be made for $e = 0$,
- equilibria $\bar{x}_3$ and $\bar{x}_6$ are unstable,
- equilibrium $\bar{x}_8$ is asymptotically stable for $e > 0$ and no conclusion can be made for $e = 0$.

An important case to study remains: the stability properties of $\bar{x}_1$ in the case that $e = 0$. While we cannot deduce (asymptotic) stability from $J(\bar{x}_1)$, the consideration of $J(\bar{x}_4)$ for $e \searrow 0$ suggests that the origin should indeed be asymptotically stable even for $e = 0$. We will come back to this question in Section 3.4 after considering some sample trajectories for time varying inputs in the next section, so as to gain more insight into the system.

3.3 Sample trajectories

To demonstrate the nonlinear effects of the model, let us consider the system response to two marginally different, piece-wise constant inputs

$$e_1(t) = \begin{cases} 0 & \text{if } t < 20 \\ 0.8 & \text{if } 20 \le t < 24 \\ 0.01 & \text{if } 24 \le t < 60 \\ 0.4 & \text{if } 60 \le t < 70 \\ 0.1 & \text{if } t \ge 70 \end{cases}$$

and

$$e_2(t) = \begin{cases} 0 & \text{if } t < 30 \\ 0.8 & \text{if } 30 \le t < 34 \\ 0.01 & \text{if } 34 \le t < 60 \\ 0.4 & \text{if } 60 \le t < 70 \\ 0.1 & \text{if } t \ge 70 \end{cases}.$$

The difference is the occurrence of the first step of the external stimulus, which happens at $t = 20$ and $t = 30$, respectively, but lasts the same duration, and the remainder of what might be thought of as environmental exposure is the same. Let us assume that for both inputs the system starts at $x^0 = (0.01, 0.01)^T$.

We find that the individual subject to $e_1(t)$ ultimately seems to recover from the external stimuli, or at least we can conclude that the level of the states remains bounded away from the point $(1, 1)$ (because we see a monotonic decrease in both states and the system is monotone).

If the first stimulus happens 10 time units later, though, we see that our individual, if instead subjected to $e_2(t)$ does not recover, and propels into a severe depression, as both states asymptotically reach 1.

Investigating the time evolution, we see two qualitatively rather similar evolutions up to the end of the second stimulus. Keeping in mind that the depression state is the only thing that can be somewhat measured (even though not with great accuracy and resolution through questionnaires), the obvious conclusion must be that timing is important for the overall outcome for the patient, as is the elusive memory state.

Another insight compatible with anecdotal evidence is that removal of external stimuli alone may not lead to recovery of an individual. For this consider the input

$$e_3(t) = \begin{cases} 0 & \text{if } t < 20 \\ 0.5 & \text{if } 20 \le t < 33 \\ 0 & \text{if } 33 \le t < 200 \\ 0.1 & \text{if } 200 \le t < 210 \\ 0 & \text{if } t \ge 210 \end{cases}$$

corresponding to a moderately severe life event at $t = 20$ and another mild one at $t = 200$. Neither event is particularly long, but the individual has a prior memory of depression with $m_0 = 0.5$ (and $s_0 = 0.01$ as before). If left without external stimulus, the memory would fade away to zero over time. But due to the first stimulus commencing at $t = 20$, the memory value is increased, and increased too far, so that it does not recover after the stimulus has ceased. Now a later and in comparison small, secondary stimulus at $t = 200$ is sufficient to push the individual into the a severe depression from which there is no recovery, even after all stimuli are over.

3.4 Consequences of monotonicity

Much of the following analysis hinges on the following fact, which can be gathered from Chapter 1 of Smith (1995) utilizing the fact that our system (3) is cooperative. By ordering we mean the component-wise ordering, i.e., for vectors x, y we have $x \ge y$ if for each component i we have $x_i \ge y_i$.

Theorem 1. (e.g., Smith (1995)). Let $x^0 \ge y^0$ be ordered initial conditions, and let locally integrable input signals e_1, e_2 be ordered point-wise, i.e., $e_1(t) \ge e_2(t)$ for all t. Denote by $x(t)$ the solution to the initial value problem $\dot{x} = f(x, e_1(t))$ with $x(0) = x^0$, and similarly $y(t)$ for the initial condition y^0 and input $e_2(t)$. Then the solutions remain ordered for all times, that is $x(t) \ge y(t)$ for all $t \ge 0$.

The first consequence is a stability result for the origin.

Corollary 2. If there is an $e^* < 1 - 2\sqrt{5}/5$ such that $e(t) < e^*$ for all $t \ge 0$ and $x^0 \le \bar{x}_4(e^*) := (\frac{e^*+\sqrt{5\,(e^*)^2-10\,e^*+1}-1}{2\,(e^*-2)}, -\frac{3\,e^*+\sqrt{5\,(e^*)^2-10\,e^*+1}-1}{2\,(e^*+1)})^T$, then the solution to the initial value problem $\dot{x} = f(x, e(t))$, $x(0) = x^0$, satisfies $x(t) \le \bar{x}_4(e^*)$ for all $t \ge 0$.

Since the order interval $[0, \bar{x}_4(e^*)] := \{x \in \mathbb{R}^2 \colon (0,0)^T \le x \le \bar{x}_4(e^*)\}$ defines an arbitrarily small neighborhood of the origin with e^* arbitrarily close to zero, we have the more specific result:

Corollary 3. If $e \equiv 0$ then the origin is stable with respect to (3).

The corollary states that if an individual with mild depression symptoms and very minor memory of such symptoms

experiences no further external stimuli, then neither the depression nor memory of it does increase.

A different question is whether both of the states actually dissipate to zero in this scenario. For that we need a different tool.

Asymptotic stability analysis of the origin in the case $e = 0$. In this case we can easily generate a specific trajectory for $e \equiv 0$ that converges to the origin. This trajectory comes from the initial condition $x^0 = \bar{x}_4(0.1)$. And because this trajectory converges to the origin as $t \to \infty$, we conclude with the help of Theorem 1 that *all* trajectories commencing in a neighborhood of the origin must converge to the origin, as they are dominated by one that does converge. We summarize this observation as follows.

Corollary 4. For $e \equiv 0$ the origin is asymptotically stable.

This result supports the intuition that "time heals all wounds" at least as long as the wound is a) small enough and b) any external stimuli causing further damage are completely removed.

4. CONCLUSION AND OUTLOOK

Our analysis is preliminary at this stage, but it is able to recover with a basic mathematical model a range of features that seem compatible with anecdotal evidence observed in individuals suffering from depression.

The model follows the paradigm to be as simple as possible but not simpler. It supports observations that a patient should continue to take their medication despite already feeling better—as the internal memory state may not have recovered yet and as this may take a substantially longer amount of time. It also supports scenarios where a patient is catapulted back into a severe depression by seemingly small trigger events that leave healthy individuals unaffected.

At the same time, the modeling process used here has made no attempt to seek grounding in more elaborate models of schema therapy or causal networks of signaling pathways in the brain as they are studied in neuroscience. The objective here was merely to propose a model that captures essential qualitative behavior, while being tractable by standard methods and concepts of the systems theory community, and we argue that this objective has been achieved.

A number of extensions are of course possible. On the mathematical side, more analysis can be done, for example by computing regions of attraction, by considering stability notions such as input-to-state stability, by designing observers for the memory state using only measurements of inputs and possibly quantized measurements of the symptom state. At the expense of simplicity, the model could further be augmented to account for treatment options such as medication, or to accommodate multiple symptoms and memory features or a resilience concept.

ACKNOWLEDGEMENTS

The author thanks practicing psychiatrist Irosh Fernando for discussions that inspired this paper.

REFERENCES

Beck, A.T., Steer, R.A., Brown, G.K., et al. (1996). Manual for the Beck depression inventory-II.

Beck, A.T., Steer, R.A., and Pompili, M. (1988). *BHS, Beck hopelessness scale: manual*. Psychological corporation San Antonio, TX.

Byrum, C.E., Ahearn, E.P., and Krishnan, K.R.R. (1999). A neuroanatomic model for depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 23(2), 175–193.

Disner, S.G., Beevers, C.G., Haigh, E.A.P., and Beck, A.T. (2011). Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience*, 12(8), 467–477.

Smith, H.L. (1995). *Monotone dynamical systems*, volume 41 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

World Health Organization (2021). Depression. URL `https://www.who.int/news-room/fact-sheets/detail/depression`.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Signed Tropicalizations of Convex Semialgebraic Sets ⋆

Mateusz Skomra *

* *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France (e-mail: mateusz.skomra@laas.fr).*

Abstract: We study the signed valuations of convex semialgebraic sets defined over non-Archimedean fields. This is motivated by the efforts to understand the structure of semialgebraic sets that arise in convex optimization, such as the spectrahedra and the hyperbolicity cones. We give a full characterization of regular sets that are obtained as signed tropicalizations of convex semialgebraic sets, and we prove that the signed tropicalizations of hyperbolicity cones have a more restrictive structure. To obtain our results, we combine two recent advances in the area of tropical geometry: the study of signed valuations of general semialgebraic sets and the separation theorems for signed tropical convexities.

Keywords: Convex algebraic geometry, Tropical geometry, Hyperbolicity cones, Semialgebraic sets, Non-Archimedean fields

AMS Subject Classification: 90C25, 14T05, 14P10

1. INTRODUCTION

Convex semialgebraic sets arise naturally in convex optimization problems such as the semidefinite programming or the hyperbolic programming. As a result, questions motivated by optimization problems, such as the study of the expressivity of semidefinite programming, lead to questions in real algebraic geometry (e.g., study of classes of sets that are representable by linear matrix inequalities). We refer to Blekherman et al. (2013) for more information about the interactions between these disciplines. In this work, we apply techniques from tropical geometry to study convex semialgebraic sets.

One of the research directions in tropical geometry is to analyze the (semi)algebraic sets defined over non-Archimedean fields with the help of the valuation map. This idea was first applied to sets arising in convex optimization by Develin and Yu (2007), who studied the tropicalizations of polyhedra. This inspired numerous other works on the tropicalizations of polyhedra, see, e.g., Allamigeon et al. (2015, 2018); Allamigeon and Katz (2017); Joswig and Smith (2018). The study of tropical polyhedra was also extended to more general semialgebraic sets, such as spectrahedra (Yu, 2015; Allamigeon et al., 2020), hyperbolicity cones in dimension 3 (Le Texier, 2021), convex semialgebraic sets (Allamigeon et al., 2019), and arbitrary semialgebraic sets (Alessandrini, 2013; Allamigeon et al., 2020; Jell et al., 2022). A recent idea proposed by Jell et al. (2022) is to study the tropicalizations of semialgebraic sets using a *signed* valuation map. In other recent development, Loho and Végh (2020) and Loho and Skomra (2022b) started a systematic study of tropical convexities in the signed setting and proved new tropical analogues of the hyperplane separation theorem. In this work, we combine both ideas by studying the *signed* valuations of *convex* semialgebraic sets. Our main goal is to generalize the known results about tropical polyhedra and spectrahedra to the tropicalizations of hyperbolicity cones in arbitrary dimension. We present the first results in this direction, by characterizing the regular sets arising as tropicalizations of convex semialgebraic sets, and by showing that the tropicalizations of hyperbolicity cones have a more restricted, "tropically quadratic" structure.

⋆ This work has benefited from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions, grant agreement 813211 (POEMA).

2. PRELIMINARIES

2.1 Generalized Puiseux series

In this work, we denote by $\mathbb{K}$ the field of *absolutely convergent generalized real Puiseux series*, i.e., series of the form

$$\boldsymbol{x} = \boldsymbol{x}(t) = c_1 t^{\alpha_1} + c_2 t^{\alpha_2} + \dots , \tag{1}$$

where both the coefficients c_i and the exponents α_i are real numbers. We further suppose that the sequence $(\alpha_i)_{i\geqslant 1}$ is strictly decreasing and either finite or unbounded and that the series $\boldsymbol{x}(t)$ is absolutely convergent for all sufficiently large $t > 0$. It is known that $\mathbb{K}$ is a real closed field (van den Dries and Speissegger, 1998). In particular, $\mathbb{K}$ is ordered by putting $\boldsymbol{x} \geqslant 0 \iff c_1 \geqslant 0$. We state our results for $\mathbb{K}$, but we note that a quantifier elimination argument discussed in Allamigeon et al. (2020) allows to transfer our main theorems from $\mathbb{K}$ to any real closed field equipped with a nontrivial and convex valuation whose value group is $\mathbb{R}$.

As a non-Archimedean field, $\mathbb{K}$ has a valuation function $\mathsf{val}\colon \mathbb{K} \to \mathbb{R} \cup \{-\infty\}$ that maps a series of the form (1) to its leading exponent, $\mathsf{val}(\boldsymbol{x}) = \alpha_1$, with $\mathsf{val}(0) = -\infty$. We note that the usual convention in the theory of valued fields would be to define the valuation as $-\alpha_1$ rather than α_1. We use the opposite convention for the sake of coherence

with the max-plus tropical semiring introduced below. One can easily adapt our results to the other setting. We also use an extended valuation map that keeps track not only of the leading exponent of the series but also of its sign.

Definition 1. We define the *signed valuation* $\mathsf{sval}\colon \mathbb{K} \to (\{-1,1\}\times\mathbb{R})\cup\{-\infty\}$ by setting $\mathsf{sval}(\boldsymbol{x}) = \big(\operatorname{sign}(\boldsymbol{x}), \mathsf{val}(\boldsymbol{x})\big)$, with the convention that $\mathsf{sval}(0) = -\infty$.

We extend the definitions of val and sval to vectors by applying them coordinatewise. A *signed tropicalization* of a semialgebraic set $\boldsymbol{X} \subset \mathbb{K}^n$ is then given by $\mathsf{sval}(\boldsymbol{X})$. An alternative viewpoint follows from the work of Alessandrini (2013), who showed that if $\boldsymbol{X} \subset \mathbb{K}^n_{>0}$, then $\mathsf{val}(\boldsymbol{X})$ coincides with the "log-limit" of sets $\boldsymbol{X}(t) \subset \mathbb{R}^n_{>0}$ obtained by fixing the parameter t, $\mathsf{val}(\boldsymbol{X}) = \lim_{t\to\infty} \log_t\big(\boldsymbol{X}(t)\big)$. In this way, if $\boldsymbol{X} \subset \mathbb{K}^n$, then $\mathsf{sval}(\boldsymbol{X})$ is obtained by "gluing" the log-limits given by all sign patterns.

2.2 Signed tropical numbers

Tropicalizations of semialgebraic sets are studied with the help of an algebraic structure known as the tropical semiring, see Maclagan and Sturmfels (2015); Joswig (2021) for more information. The *tropical (max-plus) semiring* is defined as $\mathbb{T} = (\mathbb{R}\cup\{-\infty\}, \oplus, \odot)$, where $a\oplus b = \max\{a,b\}$ and $a\odot b = a+b$. In order to replace valuation by the signed valuation, we extend $\mathbb{T}$ to signed tropical numbers. The set of signed tropical numbers is $\mathbb{T}_\pm = (\{-1,1\}\times\mathbb{R})\cup\{-\infty\}$. By convention, we denote the numbers of the form $(1,a)$ by a and call them *positive*. We also denote the numbers of the form $(-1,a)$ by $\ominus a$ and call them *negative*. Then, the set $\mathbb{T}_\pm$ is ordered by mimicking the order of the real line, with $-\infty$ having the role of zero, so that

$$\ominus 2 < \ominus 1 < \ominus(-1) < -\infty < (-1) < 1 < 2\,.$$

We denote by $[a,b]$ the interval from a to b in $\mathbb{T}_\pm$ and we embed the set $\mathbb{T}$ in $\mathbb{T}_\pm$ by identifying it with the set of signed tropical numbers that are not smaller than $-\infty$. We equip $\mathbb{T}_\pm$ with the topology induced by the order and use the product topology on $\mathbb{T}^n_\pm$. We also use a sign function $\operatorname{tsign}\colon \mathbb{T}_\pm \to \{-1,0,1\}$ that gives the sign of a signed tropical number, $\operatorname{tsign}(-\infty) = 0$, and an absolute value function $|\cdot|\colon \mathbb{T}_\pm \to \mathbb{T}$ that acts by forgetting the sign of a tropical number. In this way, the function $\phi\colon \mathbb{T}_\pm \to \mathbb{R}$ defined as $\phi(x) = \operatorname{tsign}(x)\exp(|x|)$ is an order-preserving homeomorphism. Thus, $\mathbb{T}^n_\pm$ and $\mathbb{R}^n$ are homeomorphic.

In order to equip $\mathbb{T}_\pm$ with an algebraic structure, we first define the multiplication on $\mathbb{T}_\pm$ (still denoted by $\odot$) as $a\odot b = -\infty$ if $-\infty \in \{a,b\}$ and

$$a\odot b = \big(\operatorname{tsign}(a)\operatorname{tsign}(b), |a|+|b|\big)$$

otherwise. As an example, $(\ominus 2)\odot 3 = \ominus 5$, $(\ominus 3)\odot(\ominus 4) = 7$. Defining the addition is more problematic as there is no way to define an addition that extends $\oplus$ and turns $\mathbb{T}_\pm$ into a semiring. As a way to overcome this difficulty, we equip $\mathbb{T}_\pm$ with a multivalued addition $\boxplus\colon \mathbb{T}_\pm \to 2^{\mathbb{T}_\pm}$ defined as

$$a \boxplus b = \begin{cases} a & \text{if } |a| > |b| \text{ or } a = b,\\ b & \text{if } |b| > |a|,\\ [\ominus|a|, |a|] & \text{otherwise.}\end{cases}$$

In this way $(\ominus 2)\boxplus 3 = 3$, but $(\ominus 2)\boxplus 2 = [\ominus 2, 2]$. We extend the addition to vectors by applying it coordinatewise and to sets $A, B \subset \mathbb{T}_\pm$ by putting $A\boxplus B = \cup\{a\boxplus b\colon a\in A, b\in B\}$. These operations turn $(\mathbb{T}_\pm, \boxplus, \odot)$ into a *hyperfield*, see Baker and Bowler (2019) for more information. The following lemma gives a link between the signed valuation and the hyperfield operations.

Lemma 2. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{K}$ we have $\mathsf{sval}(\boldsymbol{xy}) = \mathsf{sval}(\boldsymbol{x}) \odot \mathsf{sval}(\boldsymbol{y})$ and $\mathsf{sval}(\boldsymbol{x}+\boldsymbol{y}) \in \mathsf{sval}(\boldsymbol{x}) \boxplus \mathsf{sval}(\boldsymbol{y})$.

2.3 Tropical polynomials

A *(signed) tropical polynomial* $P(x) \in \mathbb{T}_\pm[x]$ is an expression of the form

$$\bigboxplus_{\alpha\in\Lambda} c_\alpha \odot x_1^{\odot\alpha_1} \odot \ldots \odot x_n^{\odot\alpha_n}\,,$$

where $c_\alpha \in \mathbb{T}_\pm$. A tropical polynomial defines a multivalued function $P\colon \mathbb{T}^n_\pm \to 2^{\mathbb{T}_\pm}$. One can check that if we evaluate P on a point $x \in \mathbb{T}^n_\pm$, then the result is either a singleton in $\mathbb{T}_\pm$ or an interval of the form $[\ominus a, a]$ for some $a > -\infty$. We define the *(signed) tropical hypersurface* of P as the set $\{x \in \mathbb{T}^n_\pm\colon -\infty \in P(x)\}$. Furthermore, given a polynomial $\boldsymbol{P} \in \mathbb{K}[x]$ of the form $\boldsymbol{P}(\boldsymbol{x}) = \sum_{\alpha\in\Lambda} \boldsymbol{c}_\alpha \boldsymbol{x}_1^{\alpha_1}\ldots\boldsymbol{x}_n^{\alpha_n}$, we define its *formal tropicalization* as $\operatorname{trop}(\boldsymbol{P}) = \bigboxplus_{\alpha\in\Lambda} \mathsf{sval}(\boldsymbol{c}_\alpha) \odot x_1^{\odot\alpha_1} \odot \ldots \odot x_n^{\odot\alpha_n}$. The next lemma follows from Lemma 2 and links the classical and tropical hypersurfaces.

Lemma 3. Let $\boldsymbol{P} \in \mathbb{K}[x]$ be a polynomial. Then, the set

$$\mathsf{sval}\big(\{\boldsymbol{x} \in \mathbb{K}^n\colon \boldsymbol{P}(\boldsymbol{x}) = 0\}\big)$$

is included in the tropical hypersurface of $\operatorname{trop}(\boldsymbol{P})$. In general, this inclusion may be strict.

Example 4. Consider the bivariate polynomial $\boldsymbol{P} \in \mathbb{K}[x]$ defined as $\boldsymbol{P}(\boldsymbol{x}) = (\boldsymbol{x}_1 - 1 - t^{-1})^4 + \boldsymbol{x}_2^4 - 1$. By opening the parentheses, one can see that the formal tropicalization of $\boldsymbol{P}$ is given by $x_2^{\odot 4} \boxplus x_1^{\odot 4} \boxplus (\ominus x_1^{\odot 3}) \boxplus x_1^{\odot 2} \boxplus (\ominus x_1) \boxplus (-1)$. The tropical hypersurface of $\operatorname{trop}(\boldsymbol{P})$ is the boundary of the set depicted in Figure 1. In this case, the inclusion from Lemma 3 is satisfied as an equality.

2.4 Tropical convexity

In the unsigned case, we say that a subset $X \subset \mathbb{T}^n$ is *tropically convex* if for every $x, y \in X$ and for all $\lambda, \mu \in \mathbb{T}$ such that $\lambda\oplus\mu = 0$ we have $(\lambda\odot x)\oplus(\mu\odot y) \in X$. We note that this mimics the definition of convexity over $\mathbb{R}^n$, since 0 is the neutral element of $\odot$ and the weights λ, μ are always "nonnegative", as they satisfy $\lambda, \mu \geqslant -\infty$. The following lemma relates the classical and tropical convexity, see, e.g., Develin and Sturmfels (2004), Develin and Yu (2007), Allamigeon et al. (2019) for more information.

Lemma 5. If $\boldsymbol{X} \subset \mathbb{K}^n$ is convex, then $\mathsf{val}(\boldsymbol{X}) \subset \mathbb{T}^n$ is tropically convex.

The following extension of tropical convexity to $\mathbb{T}_\pm$ was introduced by Loho and Végh (2020).

Definition 6. We say that a set $X \subset \mathbb{T}^n_\pm$ is *TO-convex* if for every $x, y \in X$ and for all $\lambda, \mu \in \mathbb{T}_\pm$ such that $\lambda, \mu \geqslant -\infty$ and $\lambda \boxplus \mu = 0$ we have $(\lambda\odot x)\boxplus(\mu\odot y) \subset X$.

The drawback of multivalued addition is that the TO-convexity is a rather strong property. In particular, Lemma 5 does not generalize to the signed setting (this requires to use a weaker notion of TC-convexity). Nevertheless, we still have a weaker property.

Lemma 7. (Loho and Skomra (2022a)). If $\boldsymbol{X} \subset \mathbb{K}^n$ is convex, then the interior of $\mathsf{sval}(\boldsymbol{X}) \subset \mathbb{T}^n_\pm$ is TO-convex.

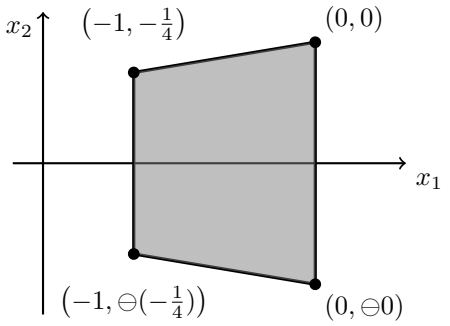


Fig. 1. Signed valuation of a convex semialgebraic set.

Example 8. The *TV screen set* is a convex set in $\mathbb{R}^2$ defined as $\{x \in \mathbb{R}^2 \colon x_1^4 + x_2^4 \leqslant 1\}$. The convexity of the TV screen set implies that the set $\boldsymbol{S}_1 = \{\boldsymbol{x} \in \mathbb{K}^2 \colon (\boldsymbol{x}_1 - 1 - t^{-1})^4 + \boldsymbol{x}_2^4 \leqslant 1\}$ is also convex. Figure 1 depicts the set $\mathsf{sval}(\boldsymbol{S}_1)$. We note that this set is TO-convex.

Example 9. The *bean curve* is the curve defined by the polynomial $p(x) = x_1(x_1^2 + x_2^2) - x_1^4 - x_1^2x_2^2 - x_2^4$. The region $\{x \in \mathbb{R}^2 \colon p(x) \geqslant 0\}$ is convex. This implies that the set $\boldsymbol{S}_2 = \{\boldsymbol{x} \in \mathbb{K}^2 \colon \boldsymbol{x}_1 \geqslant t^{-1}, p(1 - \boldsymbol{x}_1, \boldsymbol{x}_2) \geqslant 0\}$ is also convex. Its signed valuation is the same as in the previous example, $\mathsf{sval}(\boldsymbol{S}_1) = \mathsf{sval}(\boldsymbol{S}_2)$.

3. MAIN RESULTS

We now state our main results concerning the signed valuations of convex semialgebraic sets. Some of these results were obtained in collaboration with Georg Loho.

For the purpose of this work, we say that a set $X \subset \mathbb{T}_\pm^n$ is *regular* if it is equal to the closure of its interior. Regular sets arise naturally in the study of valuations of convex sets—it is known that generic tropical polyhedra are regular (Allamigeon et al., 2015) and the same is true for tropical Metlzer spectrahedra (Allamigeon et al., 2020).

3.1 Polyhedra

The signed valuations of polyhedra are studied in Loho and Végh (2020); Loho and Skomra (2022a,b). In particular, we have the following result.

Theorem 10. (Loho and Skomra (2022a)). Let $X \subset \mathbb{T}_\pm^n$ be a regular set. Then, the following are equivalent:

(1) X is an intersection of finitely many signed tropical halfspaces.
(2) X is a signed valuation of a polyhedron.

Here, a *signed tropical halfspace* is a set of the form

$$\{x \in \mathbb{T}_\pm^n \colon P(x) \cap [-\infty, +\infty) \neq \emptyset\},$$

where $P(x)$ is an affine tropical polynomial,

$$P(x) = c_0 \boxplus (c_1 \odot x_1) \boxplus \ldots \boxplus (c_n \odot x_n).$$

For the unsigned valuation, it is known that Theorem 10 holds even without the regularity assumption, see Develin and Yu (2007); Gaubert and Katz (2011). On the other hand, some assumption is necessary in the signed case. Indeed, an intersection of finitely many signed tropical halfspaces does not need to be connected, whereas a signed valuation of a convex set is always connected. It is an open question if Theorem 10 is true if we replace the assumption "X is regular" by "X is connected".

3.2 General convex semialgebraic sets

In order to characterize the regular sets that arise as images of convex semialgebraic sets, we need the following definition.

Definition 11. Let $X \subset \mathbb{T}_\pm^n$ be any set and let $\sigma \in \{-1, 1\}^n$ be a vector of signs. Then, the *maximal stratum of X given by σ* is the subset of $\mathbb{R}^n$ defined as

$$\mathrm{str}(\sigma, X) = \{(|x_1|, \ldots, |x_n|) \colon x \in X \wedge \forall i, \mathrm{tsign}(x_i) = \sigma_i\}.$$

We also say that a closed subset of $\mathbb{R}^n$ is *semilinear* if it is a union of polyhedra of the form $\{x \colon Ax \geqslant b\}$, where $b \in \mathbb{R}^m$ and $A \in \mathbb{Q}^{m \times n}$ is a rational matrix. The following theorem is our main result about tropicalizations of convex semialgebraic sets.

Theorem 12. Suppose that the set $X \subset \mathbb{T}_\pm^n$ is regular. Then, the following are equivalent:

(1) X has a TO-convex interior and semilinear maximal strata.
(2) X is a signed valuation of a convex semialgebraic set.

For the unsigned valuation, a full characterization that does not require regularity was given by Allamigeon et al. (2019). As in the case of polyhedra, it is an open question to obtain a generalization of Theorem 12 that does not require regularity.

The next corollary gives more insight into the structure of signed valuations of convex cones.

Corollary 13. Suppose that $\boldsymbol{X} \subset \mathbb{K}^n$ is a convex semialgebraic cone such that $\mathsf{sval}(\boldsymbol{X})$ is regular. Let $\mathcal{W}$ be any nonempty maximal stratum of $\mathsf{sval}(\boldsymbol{X})$. Then, $\mathcal{W}$ is a support of a pure polyhedral complex of dimension n. Furthermore, if F is an $(n-1)$-dimensional face of this complex, then the affine space spanned by F is of the form

$$\{x \in \mathbb{R}^n \colon \lambda + x_k = p^T x\}$$

for some $k \in [n]$, $\lambda \in \mathbb{R}$, and $p \in \mathbb{Q}^n_{\geqslant 0}$ that satisfies $p_k = 0$ and $\sum_{i=1}^n p_i = 1$.

Example 14. Consider the cone $\boldsymbol{S}_3 = \{\boldsymbol{x} \in \mathbb{K}^3 \colon \left(\boldsymbol{x}_1 - (1 + t^{-1})\boldsymbol{x}_3\right)^4 + \boldsymbol{x}_2^4 - \boldsymbol{x}_3^4 \leqslant 0, \boldsymbol{x}_3 \geqslant 0\}$, which is a homogenized version of the set $\boldsymbol{S}_1$ from Example 8. Its signed valuation $S = \mathsf{sval}(\boldsymbol{S}_3)$ is therefore a (tropically) homogenized version of the set depicted in Figure 1. If we fix $\sigma = (1, 1, 1)$, then $\mathrm{str}(\sigma, S)$ is a polyhedral complex with three faces of dimension 2. These faces are included in affine spaces given by $x_3 = x_1$, $x_3 - 1 = x_1$, and

$$x_2 = \frac{1}{4}x_1 + \frac{3}{4}x_3. \tag{2}$$

3.3 Hyperbolicity cones

Let us recall that a homogeneous polynomial $\boldsymbol{P} \in \mathbb{K}[x]$ is *hyperbolic with respect to* $\boldsymbol{e} \in \mathbb{K}^n$ if $\boldsymbol{P}(\boldsymbol{e}) > 0$ and, for all $\boldsymbol{x} \in \mathbb{K}^n$, all the roots of the univariate polynomial $\boldsymbol{\lambda} \mapsto \boldsymbol{P}(\boldsymbol{e} - \boldsymbol{\lambda}\boldsymbol{x})$ belong to $\mathbb{K}$. If $\boldsymbol{P}$ is hyperbolic with respect to $\boldsymbol{e}$, then the set

$$\{\boldsymbol{x} \in \mathbb{K}^n \colon \boldsymbol{\lambda} \mapsto \boldsymbol{P}(\boldsymbol{e} - \boldsymbol{\lambda}\boldsymbol{x}) \text{ has only nonnegative roots}\}$$

is called its *hyperbolicity cone*. The following lemmas summarize basic properties of hyperbolicity cones. They are stated over $\mathbb{R}$ in Renegar (2006), but they hold in all real closed fields.

Lemma 15. Hyperbolicity cones are convex.

Lemma 16. The hyperbolicity cone of $\boldsymbol{P}$ with respect to $\boldsymbol{e}$ is equal to the closure of the connected component of the set $\{\boldsymbol{x} \in \mathbb{K}^n \colon \boldsymbol{P}(\boldsymbol{x}) > 0\}$ that contains $\boldsymbol{e}$.

Our first result gives a tropical analogue of Lemma 16.

Proposition 17. Suppose that $\boldsymbol{X} \subset \mathbb{K}^n$ is a hyperbolicity cone of $\boldsymbol{P}$ with respect to some point. Furthermore, suppose that $\mathsf{sval}(\boldsymbol{X})$ is regular and let e be any point in the interior of $\mathsf{sval}(\boldsymbol{X})$. Then, $\mathsf{sval}(\boldsymbol{X})$ is equal to the closure of the connected component of the set

$$\{x \in \mathbb{T}_{\pm}^n \colon \operatorname{trop}(\boldsymbol{P})(x) \text{ is a positive singleton}\}$$

that contains e.

We note that the boundary of the connected component mentioned in Proposition 17 belongs to the tropical hypersurface of $\operatorname{trop}(\boldsymbol{P})$. Proposition 17 implies that this boundary is a subset of $\mathsf{sval}\big(\{\boldsymbol{x} \in \mathbb{K}^n \colon \boldsymbol{P}(\boldsymbol{x}) = 0\}\big)$.

Our final result implies that signed valuations of hyperbolicity cones have more restricted structure than signed valuations of general convex semialgebraic sets.

Theorem 18. Suppose that $\boldsymbol{X} \subset \mathbb{K}^n$ is a hyperbolicity cone such that $\mathsf{sval}(\boldsymbol{X})$ is regular. Let $\mathcal{W}$ be a nonempty maximal stratum of $\mathsf{sval}(\boldsymbol{X})$ and let F be an $(n-1)$-dimensional face of a polyhedral complex with support $\mathcal{W}$. Then, the affine space spanned by F is of the form

$$\{x \in \mathbb{R}^n \colon \lambda + x_k = p^T x\}, \tag{3}$$

for some $k \in [n]$, $\lambda \in \mathbb{R}$, and $p \in \{0, \frac{1}{2}, 1\}^n$ that satisfies $p_k = 0$ and $\sum_{i=1}^n p_i = 1$. In particular, p has either one or two nonzero coefficients.

Example 19. Equation (2) implies that the set $\boldsymbol{S}_3$ from Example 14 does not satisfy the condition of Theorem 18. In particular, $\boldsymbol{S}_3$ is not a hyperbolicity cone. By the discussion in Examples 8 and 9, we recover the known fact that neither the TV screen set nor the set bounded by the bean curve are spectrahedral, see Helton and Vinnikov (2007); Henrion (2010) for more discussion.

Informally, Theorem 18 can be interpreted by saying that signed valuations of hyperbolicity cones are "tropically quadratic" since the hyperplanes of the form (3) are given by tropical polynomials of degree 2. In this sense, Theorem 18 generalizes the result of Allamigeon et al. (2020) who showed that generic tropical Metzler spectrahedral cones are described by systems of tropical quadratic inequalities. We note however that Theorem 18 is only a local result: it states that every face separately can be described by a tropically quadratic polynomial, but does not give a global characterization of $\mathsf{sval}(\boldsymbol{X})$. I particular, the following is the main open question about the tropicalizations of hyperbolicity cones. This question is the tropical analogue of the generalized Lax conjecture, see Amini and Brändén (2018) for more information.

Problem 20. Suppose that $X \subset \mathbb{T}_{\pm}^n$ is a signed valuation of a hyperbolicity cone. Does this imply that X is a signed valuation of a spectrahedron?

We note that if the answer to Problem 20 is negative and a counterexample is obtained from a linear transformation of a real polynomial (like in Example 14), then we would obtain a negative answer to the generalized Lax conjecture over the real numbers.

REFERENCES

Alessandrini, D. (2013). Logarithmic limit sets of real semi-algebraic sets. *Adv. Geom.*, 13(1), 155–190.

Allamigeon, X., Benchimol, P., Gaubert, S., and Joswig, M. (2015). Tropicalizing the simplex algorithm. *SIAM J. Discrete Math.*, 29(2), 751–795.

Allamigeon, X., Benchimol, P., Gaubert, S., and Joswig, M. (2018). Log-barrier interior point methods are not strongly polynomial. *SIAM J. Appl. Algebra Geom.*, 2(1), 140–178.

Allamigeon, X., Gaubert, S., and Skomra, M. (2019). The tropical analogue of the Helton–Nie conjecture is true. *J. Symbolic Comput.*, 91, 129–148.

Allamigeon, X., Gaubert, S., and Skomra, M. (2020). Tropical spectrahedra. *Discrete Comput. Geom.*, 63, 507–548.

Allamigeon, X. and Katz, R.D. (2017). Tropicalization of facets of polytopes. *Linear Algebra Appl.*, 523, 79–101.

Amini, N. and Brändén, P. (2018). Non-representable hyperbolic matroids. *Adv. Math.*, 334, 417–449.

Baker, M. and Bowler, N. (2019). Matroids over partial hyperstructures. *Adv. Math.*, 343, 821–863.

Blekherman, G., Parrilo, P.A., and Thomas, R.R. (2013). *Semidefinite Optimization and Convex Algebraic Geometry*, volume 13 of *MOS-SIAM Ser. Optim.* SIAM, Philadelphia, PA.

Develin, M. and Sturmfels, B. (2004). Tropical convexity. *Doc. Math.*, 9, 1–27 and 205–206 (erratum).

Develin, M. and Yu, J. (2007). Tropical polytopes and cellular resolutions. *Exp. Math.*, 16(3), 277–291.

Gaubert, S. and Katz, R.D. (2011). Minimal half-spaces and external representation of tropical polyhedra. *J. Algebraic Combin.*, 33(3), 325–348.

Helton, J.W. and Vinnikov, V. (2007). Linear matrix inequality representation of sets. *Comm. Pure Appl. Math.*, 60(5), 654–674.

Henrion, D. (2010). Detecting rigid convexity of bivariate polynomials. *Linear Algebra Appl.*, 432(5), 1218–1233.

Jell, P., Scheiderer, C., and Yu, J. (2022). Real tropicalization and analytification of semialgebraic sets. *Int. Math. Res. Not.*, 2022(2), 928–958.

Joswig, M. (2021). *Essentials of Tropical Combinatorics*, volume 219 of *Grad. Stud. Math.* AMS, Providence, RI.

Joswig, M. and Smith, B. (2018). Convergent Hahn series and tropical geometry of higher rank. arXiv:1809.01457.

Le Texier, C. (2021). Hyperbolic plane curves near the non-singular tropical limit. arxiv:2109.14961.

Loho, G. and Skomra, M. (2022a). Applications of signed tropical convexity. Unpublished.

Loho, G. and Skomra, M. (2022b). Signed tropical halfspaces and convexity. arxiv:2206.13919.

Loho, G. and Végh, L.A. (2020). Signed Tropical Convexity. In *Proceedings of ITCS 2020*, 24:1–24:35.

Maclagan, D. and Sturmfels, B. (2015). *Introduction to Tropical Geometry*, volume 161 of *Grad. Stud. Math.* AMS, Providence, RI.

Renegar, J. (2006). Hyperbolic programs, and their derivative relaxations. *Found. Comput. Math.*, 6(1), 59–79.

van den Dries, L. and Speissegger, P. (1998). The real field with convergent generalized power series. *Trans. Amer. Math. Soc.*, 350(11), 4377–4421.

Yu, J. (2015). Tropicalizing the positive semidefinite cone. *Proc. Amer. Math. Soc.*, 143(5), 1891–1895.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Robust Adaptive Model Predictive Control with Persistent Excitation Conditions

Xiaonan Lu *, **Mark Cannon** *

* *Department of Engineering Science, University of Oxford, OX1 3PJ, UK, wxluxiaonan@hotmail.com, mark.cannon@eng.ox.ac.uk*

Abstract: For constrained linear systems with bounded disturbances and parametric uncertainty, we propose a robust adaptive model predictive control (MPC) scheme with online parameter estimation. Constraints enforcing persistent excitation in closed loop operation are introduced to ensure asymptotic parameter convergence. The algorithm requires the online solution of a convex optimisation problem, satisfies constraints robustly, and ensures recursive feasibility and input-to-state stability. Almost sure convergence to the actual system parameters is obtained under mild conditions on stabilisability and the tightness of disturbance bounds.

Keywords: Model Predictive Control, Robust Adaptive Control, Constrained Systems, System Identfication

1. INTRODUCTION

To be effective, model predictive controllers require accurate models of the controlled system. Adaptive MPC algorithms allow model parameters to be estimated online, reducing model uncertainty without expensive or disruptive offline testing. In system identification and adaptive control, persistent excitation (PE) conditions play a key role in establishing convergence of parameter estimates (Green and Moore, 1986; Shimkin and Feuer, 1987). By incorporating constraints to ensure appropriate PE conditions, a constrained MPC strategy can impose a lower bound on the expected rate of parameter convergence. As a result, adaptive MPC has the potential to estimate system parameters while controlling the system subject to constraints. Various approaches have been proposed (Mayne, 2014), but robust, computationally tractable adaptive MPC remains an open topic under research.

Adaptive MPC strategies usually have the dual purpose of regulating the system via feedback and providing sufficient excitation for identification of the system. Different adaptive MPC approaches place varying emphasis on these two competing objectives. Some focus on robust constraint satisfaction and stability (e.g. through constraint tightening (Di Cairano, 2016), min-max cost formulations (Adetola et al., 2009; Wang et al., 2017) or tube MPC (Lorenzen et al., 2019; Lu and Cannon, 2019)) but omit persistent excitation conditions in the problem formulation. On the other hand, some approaches consider a nominal MPC problem and force the control law to be persistently exciting, but fail to ensure constraint satisfaction and closed loop system stability (Goodwin and Sin, 1984; Marafioti et al., 2014).

Other approaches aim to achieve the dual objectives of system regulation and sufficient excitation simultaneously. For example, Weiss and Di Cairano (2014) use an augmented cost function to make the resulting control law more likely to be persistently exciting, but this is not guaranteed. Tanaskovic et al. (2014) avoids imposing PE conditions by considering the discrepancy between the nominal and actual models. However, this requires a nonconvex, infinite-dimensional optimisation that can only be simplified for specific examples. Gonzalez et al. (2014) uses a dual mode control strategy that injects persistent excitation into the system whenever the state enters a target region for parameter identification. The proposed algorithm is only applicable to open-loop stable linear systems however, and the existence of the target region is example-dependent. Hernandez Vicente and Trodden (2019) propose an algorithm that satisfies a PE condition and state and input constraints recursively, but the system model cannot be adapted online. Parsi et al. (2022) explicitly predicts the effect of future model updates in order to ensure sufficiently accurate parameter estimates, but this requires the online solution of a nonconvex problem.

In addition, although the importance of Persistent Excitation conditions have been widely acknowledged in the adaptive control literature (Narendra and Annaswamy, 1987), few strategies incorporate these conditions in a convex optimisation formulation. For example, Marafioti et al. (2014) simplifies the PE condition by expressing it as a nonconvex quadratic inequality in terms of the control input. Similarly, Hernandez Vicente and Trodden (2019) demonstrate that a PE condition can be satisfied using a periodic solution computed offline, but this solution might not be optimal. Other approaches (Lu and Cannon, 2019; Lu et al., 2021) use linearisation of the PE condition around a reference trajectory to determine sufficient conditions for persistency of excitation, but are unable to ensure closed loop satisfaction of PE conditions through recursively feasible constraints.

In this work we consider linear models with parametric uncertainty and unknown bounded additive disturbances. Building on Lu et al. (2021), we propose an adaptive MPC algorithm that combines set-based parameter identification, robust regulation, and recursively feasible con-

straints. The algorithm is input-to-state stable (ISS) and ensures convergence of parameter estimates with probability 1. Using a random excitation sequence injected into a terminal control law, we derive closed loop bounds on the expectation of PE coefficients. The algorithm employs a convex online optimisation and uses a randomised check to enforce a non-convex PE condition. The algorithm is convex and computationally tractable due to its fixed-complexity polytopic tube representation.

Notation: $\mathbb{N}_{\geq 0}, \mathbb{R}_{\geq 0}$ are the non-negative integers and reals respectively, and $\mathbb{N}_{[p,q]}$ denotes $\{n \in \mathbb{N} : p \leq n \leq q\}$. The identity matrix is $\mathcal{I}$. The ith element of a vector a is $[a]_i$ and $\|a\|$ denotes the Euclidean norm. The ith row of a matrix A is $[A]_i$, and $\text{vec}(A)$ is the vector formed by stacking the columns of A. For $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$, $A\mathcal{X} = \{Ax : x \in \mathcal{X}\}$, $\mathcal{X} \oplus \mathcal{Y} = \{x + y : x \in \mathcal{X}, y \in \mathcal{Y}\}$. Expectation is $\mathbb{E}$ and $y_{k|t}$ is the k steps ahead prediction of y at time t.

2. PERSISTENT EXCITATION

The system state $x_t \in \mathbb{R}^{n_x}$, control input $u_t \in \mathbb{R}^{n_u}$ and unknown disturbance input $w_t \in \mathbb{R}^{n_w}$, satisfy

$$x_{t+1} = A(\theta^*)x_t + B(\theta^*)u_t + Fw_t. \tag{1}$$

at all times $t \in \mathbb{N}_{\geq 0}$. Matrices $A(\theta^*)$ and $B(\theta^*)$ depend on an unknown constant parameter $\theta^* \in \mathbb{R}^p$.

Assumption 1. (Additive disturbance). The disturbance sequence $\{w_t \in \mathcal{W}, t \in \mathbb{N}_{\geq 0}\}$ is independent and identically distributed (i.i.d.), $\mathbb{E}(w_t) = 0$, $\mathbb{E}(w_t w_t^\top) \succeq \epsilon_w \mathcal{I}$, $\epsilon_w > 0$, and $\mathcal{W}$ is a known convex polyhedral set.

Assumption 2. (Model parameters). (a). $A(\theta), B(\theta)$ are defined in terms of known matrices A_j, B_j, $j \in \mathbb{N}_{[0,p]}$:

$$(A(\theta), B(\theta)) = (A_0, B_0) + \sum_{i=1}^{p} (A_i, B_i)[\theta]_i, \quad \forall \theta \in \Theta_0.$$

(b). Θ_0 is a known polytopic set containing θ^*:

$$\theta^* \in \Theta_0 = \{\theta : M_\Theta \theta \leq \mu_0\} = \text{Co}\{\theta_0^{(1)}, \ldots, \theta_0^{(m)}\}.$$

(c). The pair $\big(A(\theta^*), B(\theta^*)\big)$ is reachable.
(d). $(A(\theta), B(\theta)) = (A(\theta^*), B(\theta^*))$ if and only if $\theta = \theta^*$.

2.1 Set-based parameter estimation

Parameter identification methods include recursive least squares (Heirung et al., 2017), comparison sets (Aswani et al., 2013), set membership identification (Tanaskovic et al., 2014; Lorenzen et al., 2019) and neural network training (Akpan and Hassapis, 2011). Here we use a set membership approach to enable robust satisfaction of constraints. Set-based parameter identification was proposed in (Chisci et al., 1998; Veres et al., 1999) and it was shown in Lu et al. (2021) that the estimated parameter set converges to the true parameter value with probability 1 if the associated regressor is persistently exciting (PE).

At times $t \in \mathbb{N}_{>0}$ we use observations of the state x_t to determine a set Δ_t of unfalsified model parameters. This is combined with Θ_{t-1} to construct a new parameter set estimate Θ_t. The model (1) can be rewritten as

$$x_{t+1} = \Phi(x_t, u_t)\theta^* + \phi(x_t, u_t) + Fw_t$$

where Φ_t and ϕ_t are known at time t and are defined by

$$\Phi_t = \Phi(x_t, u_t) = [A_1 x_t + B_1 u_t \ \cdots \ A_p x_t + B_p u_t] \tag{2}$$

$$\phi_t = \phi(x_t, u_t) = A_0 x_t + B_0 u_t. \tag{3}$$

Given x_t, x_{t-1}, u_{t-1} and the disturbance set $\mathcal{W}$, the unfalsified parameter set at time t is given by

$$\Delta_t = \{\theta : x_t - A(\theta)x_{t-1} - B(\theta)u_{t-1} \in F\mathcal{W}\}$$

The parameter set Θ_t may be updated using Δ_t by various methods, including minimal (Chisci et al., 1998), fixed-complexity (Lorenzen et al., 2019), and limited-complexity (Tanaskovic et al., 2014) update laws. In each case, Θ_t is non-increasing and $\Theta_t \subseteq \Theta_{t-1}$ for all $t \in \mathbb{N}_{>0}$.

For a fixed-complexity parameter set update law, the parameter set estimate Θ_t is defined as $\Theta_t = \Theta(\mu_t) = \{\theta : M_\Theta \theta \leq \mu_t\}$ where $M_\Theta \in \mathbb{R}^{r \times p}$ is an *a priori* chosen matrix and $\mu_t \in \mathbb{R}^r$ is determined so that Θ_t is the smallest set containing the intersection of Θ_{t-1} and the unfalsified sets $\Delta_{t-N_\mu+1}, \ldots, \Delta_t$,

$$\mu_t := \min_{\mu \in \mathbb{R}^r} \text{vol}\big(\Theta(\mu)\big) \ \text{ s.t. } \bigcap_{j=t-N_\mu+1}^{t} \Delta_j \cap \Theta_{t-1} \subseteq \Theta(\mu) \tag{4}$$

where $\Delta_j := \mathbb{R}^p$ for $j \leq 0$ and N_μ is the parameter update window length. Note that μ_t can be computed by solving a set of linear programs.

We briefly recap the definition of persistent excitation (PE). The regressor Φ_t in (2) is persistently exciting if there exists a horizon N_u and a scalar $\epsilon_\Phi > 0$ such that

$$\sum_{k=t}^{t+N_u-1} \Phi_k^\top \Phi_k \succeq \epsilon_\Phi \mathcal{I} \tag{5}$$

for all $t \in \mathbb{N}_{\geq 0}$. In the current work however, we define persistent excitation using the expectation condition

$$\sum_{k=t}^{t+N_u-1} \mathbb{E}\big\{\Phi_k^\top \Phi_k\big\} \succeq \epsilon_\Phi \mathcal{I}, \tag{6}$$

which is required to hold for some $\epsilon_\Phi > 0$ with non-zero probability, for all $t \in \mathbb{N}_{\geq 0}$. We refer to the interval $\mathbb{N}_{[t,t+N_u-1]}$ as a PE window.

Assumption 3. (Tight disturbance bound). For all $w^0 \in \partial\mathcal{W}$ and any $\epsilon > 0$ the disturbance sequence $\{w_0, w_1, \ldots\}$ satisfies $\Pr\big\{\|w_t - w^0\| < \epsilon\big\} \geq p_w(\epsilon)$, for all $t \in \mathbb{N}_{\geq 0}$, where $p_w(\epsilon) > 0$ whenever $\epsilon > 0$.

The following result extends Lu et al. (2021), Corollaries 2 and 3, to the case of the modified PE condition (6).

Lemma 1. Under Assumptions 1 and 3, if Φ_t satisfies the PE condition (6) with probability $p > 0$ for all t, then the minimal and fixed complexity parameter set estimates Θ_t with $N_\mu \geq N_u$ converge to $\{\theta^*\}$ with probability 1.

Assumption 1 on the disturbance sequence $\{w_t, t \in \mathbb{N}_{\geq 0}\}$ is common in practice. Assumption 3 may be more difficult to verify, but we note that this assumption can be relaxed at the expense of some residual uncertainty in the parameter set estimate (see Lu et al. (2021) for details).

2.2 Linear feedback with injected noise

Consider a feedback law with injected noise:

$$u_t = Kx_t + s_t \tag{7}$$

where s_t is a stochastic variable. To simplify notation we define $A_K(\theta) = A(\theta) + B(\theta)K$, $A_{K,i} = A_i + B_i K$, $i \in \mathbb{N}_{[0,p]}$.

Assumption 4. (Stability). For $z_t \in \mathbb{R}^{n_x}$ and $t \in \mathbb{N}_{\geq 0}$, $z_{t+1} \in \text{Co}\{A_K(\theta)z_t, \theta \in \Theta_0\}$ is quadratically stable.

Assumption 5. The sequence $\{s_t \in \mathcal{S}, t \in \mathbb{N}_{\geq 0}\}$ is i.i.d. with $\mathbb{E}(s_t) = 0$, $\mathbb{E}(s_t s_t^\top) \succeq \epsilon_s \mathcal{I}$, $\epsilon_s > 0$, s_t is independent of x_t and w_t, and $\mathcal{S}$ is a known polytopic set.

Theorem 2. If $N_u > n_x$, then under Assumptions 1, 2, 4 and 5, the regressor Φ_t in (2) of the system (1) with $u_t = Kx_t + s_t$ satisfies, for some $\epsilon_\Phi > 0$ and all $k \in \mathbb{N}_{\geq 0}$,

$$\sum_{k=t}^{t+N_u-1} \mathbb{E}\big(\Phi_k^\top \Phi_k\big) \succeq \epsilon_\Phi \mathcal{I}. \tag{8}$$

3. ADAPTIVE ROBUST MPC

The noise s_t injected into the feedback law (7) can cause poor tracking performance and may violate state and control constraints. However, a receding horizon control law incorporating injected noise can avoid these undesirable effects while exploiting the PE properties it provides. Consider a predicted control law parameterised at time t in terms of decision variables $\mathbf{v}_t = \{v_{0|t}, \dots, v_{N-1|t}\}$:

$$u_{k|t} = \begin{cases} Kx_{k|t} + v_{k|t}, & k \in \mathbb{N}_{[0,N-1]} \\ Kx_{k|t} + s_{k|t}, & k \in \mathbb{N}_{[N,N+N_u-1]} \end{cases} \tag{9}$$

where N is the prediction horizon. We assume linear state and control input constraints

$$x_{k|t} \in \mathcal{X}, \quad u_{k|t} \in \mathcal{U}, \quad \forall k \in \mathbb{N}_{[0,N+N_u-1]}, \tag{10}$$

where $\mathcal{X}$, $\mathcal{U}$ are given polytopes. To enforce these constraints we define a terminal set $\mathcal{X}_T$ satisfying

$$\mathcal{X}_T \subseteq \mathcal{X}, \quad K\mathcal{X}_T \oplus \mathcal{S} \subseteq \mathcal{U}, \tag{11}$$

$$A_K(\theta)\mathcal{X} \oplus B(\theta)\mathcal{S} \oplus F\mathcal{W} \subseteq \mathcal{X}_T \quad \forall \theta \in \Theta_t. \tag{12}$$

3.1 Tube MPC formulation

To ensure satisfaction of constraints (10) we construct a sequence of sets denoted $\mathbf{X}_t = \{\mathcal{X}_{k|t}, k \in \mathbb{N}_{[0,N+N_u-1]}\}$, satisfying, for all $\theta \in \Theta_t$ and $k \in \mathbb{N}_{[0,N-1]}$,

$$\mathcal{X}_{k|t} \subseteq \mathcal{X}, \quad K\mathcal{X}_{k|t} \oplus \{v_{k|t}\} \subseteq \mathcal{U}, \tag{13}$$

$$A_K(\theta)\mathcal{X}_{k|t} \oplus \{B(\theta)v_{k|t}\} \oplus F\mathcal{W} \subseteq \mathcal{X}_{k+1|t}. \tag{14}$$

The initial and terminal conditions are given by

$$\mathcal{X}_{0|t} = \{x_t\}, \tag{15}$$

$$\mathcal{X}_{k|t} = \mathcal{X}_T, \ k \in \mathbb{N}_{[N,N+N_u-1]} \tag{16}$$

where $\mathcal{X}_T$ satisfies (11)-(12). We consider a nominal cost defined for a given nominal parameter vector $\bar{\theta}_t \in \Theta_t$ by

$$J(x_t, \mathbf{v}_t, \bar{\theta}_t) = \sum_{k=0}^{N-1} l(\bar{x}_{k|t}, K\bar{x}_{k|t} + v_{k|t}) + V_{N|t}(\bar{x}_{N|t})$$

with $\bar{x}_{0|t} = x_t$, $\bar{x}_{k+1|t} = A_K(\bar{\theta}_t)\bar{x}_{k|t} + B(\bar{\theta}_t)v_{k|t}$, $k \in \mathbb{N}_{[0,N-1]}$. The stage cost $l(\cdot,\cdot)$ and terminal cost $V_{N|t}(\cdot)$ are assumed to be positive definite quadratic functions satisfying, for given $\bar{\theta}_t \in \Theta_0$ and all $x \in \mathbb{R}^{n_x}$,

$$V_{N|t}(x) = V_{N|t}\big(A_K(\bar{\theta}_t)x\big) + l(x, Kx). \tag{17}$$

The MPC law is determined by the solution, denoted $(\mathbf{v}_t^o, \mathbf{X}_t^o)$, of the problem of minimising $J(x_t, \mathbf{v}_t, \bar{\theta}_t)$ over $\mathbf{v}_t$ and $\mathbf{X}_t$ subject to (13)-(16) and additional constraints described in Section 3.2 to ensure the PE condition (6).

3.2 PE condition

To define a recursively feasible set of constraints, we construct a series of N_{pe} overlapping PE windows extending from the past and across the prediction horizon, where

$$N_{pe} = \begin{cases} N + t, & t < N_u - 1 \\ N + N_u - 1, & t \geq N_u - 1. \end{cases}$$

Consider the PE conditions defined at time $t \in \mathbb{N}_{>0}$ for given $\mathbf{v}_t$, $\mathbf{s}_t$, $\mathbf{X}_t$ and all $\kappa \in \mathbb{N}_{[N-N_{pe},N-1]}$ by

$$\sum_{k=\kappa}^{\kappa+N_u-1} \Phi(x_{k|t}, Kx_{k|t} + q_{k|t})^\top \Phi(x_{k|t}, Kx_{k|t} + q_{k|t}) \succeq \beta_{\kappa|t}\mathcal{I} \tag{18}$$

for all $x_{k|t} \in \mathcal{X}_{k|t}$, $k \in \mathbb{N}_{[\kappa,\kappa+N_u-1]}$, where $q_{k|t} := v_{k|t}$ if $k < N$ and $q_{k|t} := s_{k|t}$ if $k \geq N$. These conditions are nonconvex in $\mathbf{v}_t$, $\mathbf{X}_t$, and hence unsuitable as constraints in an online MPC optimisation. However, following Lu et al. (2021) we can linearise these conditions around a reference trajectory $(\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t) = \{(\hat{x}_{k|t}, \hat{u}_{k|t}), k \in \mathbb{N}_{[0,N+N_u-1]}\}$. For a given nominal parameter vector $\bar{\theta}_t$ and sequences $\hat{\mathbf{v}}_t$, $\mathbf{s}_t$, this reference trajectory is defined by

$$\hat{x}_{0|t} = x_t \tag{19a}$$

$$\hat{x}_{k+1|t} = A(\bar{\theta}_t)\hat{x}_{k|t} + B(\bar{\theta}_t)\hat{u}_{k|t}, \ k \in \mathbb{N}_{[0,N+N_u-2]} \tag{19b}$$

$$\hat{u}_{k|t} = \begin{cases} K\hat{x}_{k|t} + \hat{v}_{k|t}, & k \in \mathbb{N}_{[0,N-1]} \\ K\hat{x}_{k|t} + s_{k|t}, & k \in \mathbb{N}_{[N,N+N_u-1]}. \end{cases} \tag{19c}$$

We define the sequence $\hat{\mathbf{v}}_t$ for $t \in \mathbb{N}_{>0}$ using the solution of the MPC optimisation at time $t-1$, denoted $\mathbf{v}_{t-1}^o = \{v_{0|t-1}^o, \dots, v_{N-1|t-1}^o\}$, and $s_{N|t}$:

$$\hat{v}_{k|t} = \begin{cases} v_{k+1|t-1}^o, & k \in \mathbb{N}_{[0,N-2]} \\ s_{N-1|t}, & k = N-1. \end{cases} \tag{20}$$

Linearising (18) by neglecting quadratic terms in the decision variables $(\mathbf{v}_t, \mathbf{X}_t)$ yields a set of LMIs in $\mathbf{v}_t$, the vertices, $x_{k|t}^{(j)}$, of $\mathbf{X}_t$, and additional optimisation variables $\boldsymbol{\beta}_t' = \{\beta_{\kappa|t}', \kappa \in \mathbb{N}_{[N-N_{pe},N-1]}\}$, given for $\kappa \in \mathbb{N}_{[N-N_{pe},N-1]}$ by

$$\sum_{k=\kappa}^{-1} \Phi_{k+t}^\top \Phi_{k+t} + \sum_{k=\max\{0,\kappa\}}^{\kappa+N_u-1} M_{k|t} \succeq \beta_{\kappa|t}'\mathcal{I} \tag{21}$$

where $M_{k|t} = M_{k|t}^\top \in \mathbb{R}^{p\times p}$, $k \in \mathbb{N}_{[0,N+N_u-2]}$ satisfies

$$\begin{aligned} M_{k|t} \preceq\ & \hat{\Phi}_{k|t}^\top \hat{\Phi}_{k|t} + \hat{\Phi}_{k|t}^\top \Phi(x_{k|t}^{(j)} - \hat{x}_{k|t}, Kx_{k|t}^{(j)} + q_{k|t} - \hat{u}_{k|t}) \\ & + \Phi(x_{k|t}^{(j)} - \hat{x}_{k|t}, Kx_{k|t}^{(j)} + q_{k|t} - \hat{u}_{k|t})^\top \hat{\Phi}_{k|t}, \end{aligned}$$

for all $j \in \mathbb{N}_{[1,\nu]}$, with $\hat{\Phi}_{k|t} = \Phi(\hat{x}_{k|t}, \hat{u}_{k|t})$.

To increase the probability of the solution satisfying (6), we include (21) in the MPC optimisation at times $t > 0$ with the following constraints on $\boldsymbol{\beta}_t'$

$$\beta_{\kappa|t}' \geq \hat{\beta}_{\kappa|t}', \quad \forall \kappa \in \mathbb{N}_{[N-N_{pe},N-1]}, \tag{22}$$

where $\hat{\beta}_{\kappa|t}'$ is a lower bound on the maximum value of $\beta_{\kappa|t}'$ satisfying (21) for $(\mathbf{v}_t, \mathbf{X}_t)$ satisfying (13)-(16). Thus we determine $\hat{\boldsymbol{\beta}}_t' = \{\hat{\beta}_{\kappa|t}', \kappa \in \mathbb{N}_{[N-N_{pe},N-1]}\}$ by finding $\hat{\beta}_{\kappa|t}'$ in (22) as the solution of

$$\hat{\beta}_{\kappa|t}' := \max_{\beta_{\kappa|t}' \in \mathbb{R}} \beta_{\kappa|t}' \ \text{ s.t. } (21) \tag{23}$$

with $q_{k|t} := \hat{v}_{k|t}$ if $k < N$ and $q_{k|t} := s_{k|t}$ if $k \geq N$, and with $x_{k|t}^{(j)} := x_{k+1|t-1}^{(j)\,o}$, $\forall j \in \mathbb{N}_{[1,\nu]}$, $k > 0$, where $x_{k|t-1}^{(j)\,o}$ denotes a vertex of $\mathbf{X}_{t-1}^o$.

To impose (6) with non-zero probability, we propose a computationally undemanding check whether the solution of the online MPC optimisation using sampling:

$$\beta^s_{\kappa|t} \geq \hat{\beta}^s_{\kappa|t}, \qquad \forall \kappa \in \mathbb{N}_{[N-N_{pe},N-1]}. \tag{24}$$

Here $\beta^s_{\kappa|t}$ denotes the maximum $\beta_{\kappa|t}$ satisfying (18) with $q_{k|t} = v^o_{k|t}$, $k < N$ and $x_{k|t} \in \{x^{(1)}_{k|t}, \ldots, x^{(N_s)}_{k|t}\}$ where $x^{(i)}_{k|t} \in \mathcal{X}^o_{k|t}$ for $i \in \mathbb{N}_{[1,N_s]}$, $k \in \mathbb{N}_{[\kappa,\kappa+N_u-1]}$ are random samples, and $\hat{\beta}^s_{\kappa|t}$ is defined analogously with $q_{k|t} = \hat{v}_{k|t}$ for $k < N$ and random samples of $\mathcal{X}^o_{k+1|t-1}$, $k \in \mathbb{N}_{[\kappa,\kappa+N_u-1]}$.

Theorem 3. (Stability). Let Assumptions 1, 2 and 4 hold. Then the system (1) under Algorithm 1 is input-to-state practically stable (ISpS) (Limon et al., 2009, Def. 6) in the set of initial conditions x_0 for which $\mathcal{P}_0$ is feasible.

Theorem 4. (PE). Under Assumptions 1, 2, 4 and 5 and with $N_u > n_x$, the system (1) with the control law of Algorithm 1 satisfies the PE condition (6) with probability $p > 0$ for some $\epsilon_\Phi > 0$, for all t.

Theorem 3 implies (Limon et al., 2009; Lu et al., 2021) the existence of a $\mathcal{KL}$-function η and $\mathcal{K}$-functions ψ, ξ satisfying, for all x_0 such that $\mathcal{P}_0$ is feasible, the bound

$$\begin{aligned}\|x_t\| \leq \eta(\|x_0\|, t) + \psi\big(\max_{\tau \in \mathbb{N}_{[0,t-1]}} \|Fw_\tau + B(\theta^*)s_\tau\|\big) \\ + \xi\big(\max_{\tau \in \mathbb{N}_{[0,t-1]}} \|\bar{\theta}_\tau - \theta^*\|\big).\end{aligned} \tag{25}$$

Thus Lemma 1 and Theorems 3 and 4 imply that: (i) the parameter set estimate Θ_t converges asymptotically to $\{\theta^*\}$; (ii) system (1) is ISS under the action of Algorithm 1.

Algorithm 1 Adaptive MPC with PE constraints

At $t = 0$: Choose Θ_0 and $\mathcal{S}$. Determine K satisfying Assumption 4 and $\mathcal{X}_T$, $V_{N|0}$ satisfying (11)-(12) and (17). Define N, N_u, N_w, N_θ. Obtain $\bar{\theta}_0$ and x_0 and compute the solution, $(\mathbf{v}^o_0, \mathbf{X}^o_0)$, of the quadratic program (QP):

$$\mathcal{P}_0: \underset{\mathbf{v}_0, \mathbf{X}_0}{\text{minimise}} \; J(x_0, \mathbf{v}_0, \bar{\theta}_0) \;\; \text{s.t. (13)-(16).} \tag{26}$$

Apply the control input $u_0 = Kx_0 + v^o_{0|0}$.

At times $t = 1, 2, \ldots$:

(a). Obtain the current state x_t.

(b). Update $\bar{\theta}_t$ and Θ_t using (4) and $V_{N|t}$ via (17).

(c). Generate the noise sequence $\mathbf{s}_t$ and compute $\hat{\mathbf{x}}_t$, $\hat{\mathbf{u}}_t$, $\hat{\mathbf{v}}_t$, $\hat{\boldsymbol{\beta}}'_t$ using (19), (20) and (23).

(d). Find the solution $(\mathbf{v}^o_t, \mathbf{X}^o_t)$ of the semidefinite program

$$\begin{aligned}\mathcal{P}_{>0}: \underset{\mathbf{v_t}, \mathbf{X}_t, \boldsymbol{\beta}_t}{\text{minimise}} \; & J(x_t, \mathbf{v}_t, \bar{\theta}_t) \\ & \text{s.t. (13)-(16), (21) and (22).}\end{aligned} \tag{27}$$

(e). Generate N_s samples of $\mathcal{X}_{k|t}$ and $\mathcal{X}^o_{k+1|t-1}$ for $k \in \mathbb{N}_{[\kappa,\kappa+N_u-1]}$ and compute $\beta^s_{\kappa|t}$, $\hat{\beta}^s_{\kappa|t}$ for $\kappa \in \mathbb{N}_{[N-N_{pe},N-1]}$. If (24) is not satisfied, set $\mathbf{v}^o_t := \hat{\mathbf{v}}_t$ and $\mathbf{X}^o_t := \big\{\{x_t\}, \mathcal{X}^o_{2|t-1}, \ldots, \mathcal{X}^o_{N-1|t-1}, \mathcal{X}_T, \ldots, \mathcal{X}_T\big\}$.

(f). Apply the control input $u_t = Kx_t + v^o_{0|t}$

REFERENCES

Adetola, V., DeHaan, D., and Guay, M. (2009). Adaptive model predictive control for constrained nonlinear systems. *Sys. Con. Lett.*, 58(5), 320–326.

Akpan, V.A. and Hassapis, G.D. (2011). Nonlinear model identification and adaptive model predictive control using neural networks. *ISA Trans.*, 50(2), 177–194.

Aswani, A., Gonzalez, H., Sastry, S.S., and Tomlin, C. (2013). Provably safe and robust learning-based model predictive control. *Automatica*, 49(5), 1216–1226.

Chisci, L., Garulli, A., Vicino, A., and Zappa, G. (1998). Block recursive parallelotopic bounding in set membership identification. *Automatica*, 34(1), 15–22.

Di Cairano, S. (2016). Indirect adaptive model predictive control for linear systems with polytopic uncertainty. *American Control Conf.*, 3570–3575.

Gonzalez, A., Ferramosca, A., Bustos, G., Marchetti, J., Fiacchini, M., and Odloak, D. (2014). Model predictive control suitable for closed-loop re-identification. *Sys. Con. Lett.*, 69(1), 23–33.

Goodwin, G.C. and Sin, K.S. (1984). *Adaptive filtering prediction and control*. Prentice-Hall, Englewood Cliffs.

Green, M. and Moore, J.B. (1986). Persistence of excitation in linear systems. *Sys. Con. Lett.*, 351–360.

Heirung, T.A.N., Ydstie, B.E., and Foss, B. (2017). Dual adaptive model predictive control. *Automatica*, 80, 340–348.

Hernandez Vicente, B. and Trodden, P. (2019). Stabilizing predictive control with persistence of excitation for constrained linear systems. *Sys. Con. Lett.*, 126, 58–66.

Limon, D., Alamo, T., Raimondo, D., Muñoz de la Peña, D., Bravo, J., Ferramosca, A., and Camacho, E. (2009). Input-to-state stability: A unifying framework for robust model predictive control. In *Nonlinear Model Predictive Control*. Springer, Berlin.

Lorenzen, M., Cannon, M., and Allgöwer, F. (2019). Robust MPC with recursive model update. *Automatica*, 103, 467–471.

Lu, X., Cannon, M., and Koksal-Rivet, D. (2021). Robust adaptive model predictive control: performance and parameter estimation. *Int. J. Robust Nonlinear Control*, 31, 8703–8724.

Lu, X. and Cannon, M. (2019). Robust adaptive tube model predictive control. *American Control Conf.*, 3695–3701.

Marafioti, G., Bitmead, R.R., and Hovd, M. (2014). Persistently exciting model predictive control. *Int. J. Adaptive Control Sig. Proc.*, 28(6), 536–552.

Mayne, D.Q. (2014). Model predictive control: Recent developments and future promise. *Automatica*, 50(12), 2967–2986.

Narendra, K.S. and Annaswamy, A.M. (1987). Persistent excitation in adaptive systems. *Int. J. Control*, 45, 127–160.

Parsi, A., Iannelli, A., and Smith, R. (2022). An explicit dual control approach for constrained reference tracking of uncertain linear systems. *IEEE Transactions on Automatic Control*. Doi: 10.1109/TAC.2022.3176800.

Shimkin, N. and Feuer, A. (1987). Persistency of excitation in continuous-time systems. *Sys. Con. Lett.*, 9, 225–233.

Tanaskovic, M., Fagiano, L., Smith, R., and Morari, M. (2014). Adaptive receding horizon control for constrained MIMO systems. *Automatica*, 50, 3019–3029.

Veres, S.M., Messaoud, H., and Norton, J.P. (1999). Limited complexity model-unfalsifying adaptive tracking control. *Int. J. Control*, 72, 1417–1426.

Wang, X., Yang, L., Sun, Y., and Deng, K. (2017). Adaptive model predictive control of nonlinear systems with state-dependent uncertainties. *Int. J. Robust Nonlin. Control*, 27(17), 4138–4153.

Weiss, A. and Di Cairano, S. (2014). Robust dual control MPC with guaranteed constraint satisfaction. *IEEE Conf. Decision and Control*, 6713–6718.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Error Bounds for Locally Optimal Distributed Filters over Random Graphs

Aneel Tanwani *

* *LAAS – CNRS, University of Toulouse, Toulouse France.*
Email: aneel.tanwani@cnrs.fr

Abstract: In this extended abstract, we consider the problem of analyzing the performance of distributed filters for continuous-time linear stochastic systems under certain information constraints. We associate an undirected and connected graph with the measurements of the system, where the nodes have access to partial measurements in continuous time. Each node executes a locally optimally filter based on the available measurements. In addition, a node communicates its estimate to a neighbor at some randomly drawn discrete time instants, and these activation times of the graph edges are governed by independent Poisson counters. When a node gets some information from its neighbor, it resets its state using a convex combination of the available information. Consequently, each node implements a filtering algorithm in the form of a stochastic hybrid system. We derive bounds on expected value of error covariance for each node, and show that they converge to a common value for each node if the mean sampling rates for communication between nodes are large enough. The material covered in this extended abstract is based on the publications (Tanwani, 2021, 2022).

Keywords: Stochastic hybrid system; distributed filtering; graph theory; random communication.

1. INTRODUCTION

Modern engineering systems often involve integration of several components connected to one another to execute a complex task with efficient use of resources. Implementation of such architectures have paved way for distributed decision making and consequently, this has lead to immense research on design and analysis of distributed algorithms, see for example (DeGroot, 1974; Tsitsiklis et al., 1986). In particular, the problem of state estimation, and filtering, in dynamical systems has received particular attention (Olfati-Saber, 2007, 2009). Distributed filtering allows us to disintegrate a centralized output into several components, and then associate a filtering algorithm with each of these smaller components, see Figure 1 for a conventional layout of such architectures. In the usual operation of distributed filters, it is assumed that the sensor units, represented by the nodes in a graph, communicate the information about their own estimate to their neighbors (determined by the graph topology) at all times. In our work, however, we put constraints on the communication between these dynamic agents, which represent the individual filtering units. This makes the underlying graph time-varying and the asymptotic behavior of distributed algorithms in such cases has been studied in (Jadbabaie et al., 2003; Moreau, 2005; Cao et al., 2008). In our setup, we assume that each link in the graph is activated at random time-instants and the random process, which determines the discrete-times at which two neighbors communicate, is described by a Poisson counter. For this problem setup, we propose filtering algorithm in the form of a stochastic hybrid system. Such framework has been advocated in (Hespanha, 2014) for control problems over networks with communication constraints. Some historical developments on the use of Poisson counters for sampling process are provided in (Tanwani et al., 2018). Some recent work from the author deals with analyzing the performance of model predictive control under random sampling (Tanwani et al., 2019). The results proposed in this work build on the centralized filtering case studied in Tanwani and Yufereva (2020).

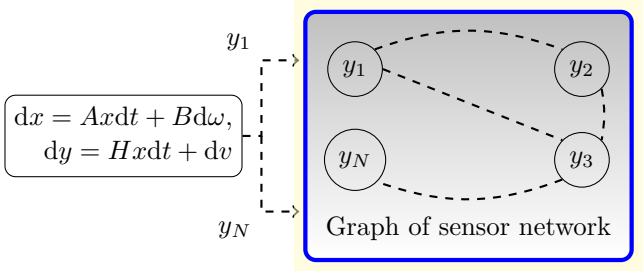


Fig. 1. Layout of distributed filters, where dashed lines represent communication at random times.

2. LOCALLY OPTIMAL DISTRIBUTED FILTERING

Let us begin with the description of the system class and the formulation of the distributed filtering problem studied in this paper. In the process, we describe the graph structure representing the interconnection between sensor nodes, and the sampling process at which the connected nodes (or the neighbors) exchange information.

⋆ This work is sponsored by the project CyPhAI, financed by ANR-JST CREST program with grant number ANR-20-JSTM-0001.

2.1 System Class

We consider dynamical systems modeled by linear stochastic differential equations of the form

$$\mathrm{d}x = Ax\,\mathrm{d}t + B\,\mathrm{d}\omega \tag{1}$$

where $(x(t))_{t\geq 0}$ is an $\mathbb{R}^n$-valued diffusion process describing the state. Let $(\Omega, \mathcal{F}, \mathsf{P})$ denote the underlying probability space. It is assumed that, for each $t \geq 0$, $(\omega(t))_{t\geq 0}$ is a zero mean $\mathbb{R}^m$-valued standard Wiener process adapted to the filtration $\mathcal{F}_t \subset \mathcal{F}$, with the property that $\mathsf{E}[\mathrm{d}\omega(t)\,\mathrm{d}\omega(t)^\top] = I_m \mathrm{d}t$, for each $t \geq 0$. The matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are taken as constant with (A, B) controllable, and the process $(\omega(t))_{t\geq 0}$ does not depend on the state. The solutions of the stochastic differential equation (1) are interpreted in the sense of Itô stochastic integral.

The centralized output measurement associated to the process (1) is of the form

$$\mathrm{d}y = Hx\,\mathrm{d}t + \mathrm{d}v \tag{2}$$

where $H \in \mathbb{R}^{p\times n}$ is a constant matrix, with (A, H) being observable, and $(v(t))_{t\geq 0}$ is a zero mean $\mathbb{R}^p$-valued standard Wiener process. The conventional filtering problem, with initial state having Gaussian distribution, deals with constructing a mean-square estimate of the state x_t, denoted by $\widehat{x}_t$ so that $\mathsf{E}[|x_t - \widehat{x}_t|^2 \,|\, (\mathrm{d}y(s))_{s\leq t}]$ is minimized. The optimal estimate which achieves this minimum value is $\mathsf{E}[x_t \,|\, (\mathrm{d}y(s))_{s\leq t}]$ and is computed recursively using a Kalman-Bucy filter. For the problem studied in this paper, it is assumed that the centralized measurements are not available and we address the filtering problem with similar assumptions on system data, but under different information constraints which are described next.

2.2 Information Structure

The measurements associated with system (1) are obtained from a set of N sensors which are distributed in their localization. Each of these sensors provides a partial measurement about the state described as,

$$\mathrm{d}y_i = H_i x\,\mathrm{d}t + \mathrm{d}v_i, \quad i = 1, \ldots, N, \tag{3}$$

where $H_i \in \mathbb{R}^{p_i\times n}$, and $\sum_{i=1}^N p_i = p$. That is, for each node, $(y_i(t))_{t\geq 0}$ describes an $\mathbb{R}^{p_i}$-valued continuous-time observation process. In the observation equation (3), $v_i(t)$ is a zero mean $\mathcal{F}_t$-adapted standard Wiener process, taking values in $\mathbb{R}^{p_i}$, and $\mathsf{E}[\mathrm{d}v_i(t)\,\mathrm{d}v_i(t)^\top] = V_i \mathrm{d}t$, with $V_i \in \mathbb{R}^{p_i\times p_i}$ assumed to be positive definite. The optimal filter which minimizes the mean square estimation error conditioned upon the information available through the measurements $\{\mathrm{d}y_i(s) | s \leq t\}$ is,

$$\mathrm{d}\widehat{x}_i(t) = A\widehat{x}_i(t)\mathrm{d}t + P_i(t)H_i^\top V_i^{-1}(\mathrm{d}y_i(t) - H_i\widehat{x}_i(t)\mathrm{d}t) \tag{4a}$$

$$\dot{P}_i = AP_i + P_iA^\top - P_iH_i^\top V_i^{-1}H_iP_i + BB^\top, \tag{4b}$$

with $\widehat{x}_i(0) = \mathsf{E}[x(0)]$, and $P_i(t)$ is exactly the error covariance $\mathsf{E}[(x_i(t) - \widehat{x}_i(t))(x_i(t) - \widehat{x}_i(t))^\top \,|\, \mathrm{d}y_i(s), s \leq t]$ if $P_i(0) = \mathsf{E}[(x_i(0) - \widehat{x}_i(0))(x_i(0) - \widehat{x}_i(0))^\top]$.

Communication Graph: The sensor nodes are connected via a graph $\mathcal{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V} = \{1, \ldots, N\}$ is the set of graph nodes, and $\mathscr{E}$ contains all the edges defined by a subset of the pairs (i, j), $i \neq j$, $i, j \in \mathscr{V}$. We assume that the graph is undirected and connected. The neighbors of a node $i \in \mathscr{V}$ are denoted by $\mathcal{N}_i$ and we adopt the convention that $i \notin \mathcal{N}_i$. The adjacency matrix $\mathcal{A} := [\alpha_{ij}] \in \{0, 1\}^{N\times N}$ of the graph, which is symmetric, provides the information about which sensor nodes can communicate with each other, that is, if $\alpha_{ij} = 1$ then sensor i and j can communicate, whereas $\alpha_{ij} = 0$ means there is no communication possible between those sensors. The degree of a node $i \in \mathscr{V}$ is defined as $|\mathcal{N}_i|$, that is, the cardinality of the set $\mathcal{N}_i$. The diagonal matrix $\mathcal{D} = [d_{ii}]$, with $d_{ii} = |\mathcal{N}_i|$ is therefore the degree matrix. We associate a Laplacian $\mathcal{L}$ with this graph, defined as, $\mathcal{L} = \mathcal{D} - \mathcal{A}$. For our purposes, the nonnegative matrix $\Pi = [\pi_{ij}] \in \mathbb{R}^{N\times N}$, defined as

$$\Pi := I_N - \epsilon\mathcal{L} \tag{5}$$

where $0 < \epsilon \leq \min_{i\in\mathscr{V}} \frac{1}{|\mathcal{N}_i|}$ plays an important role. Note that, by construction, Π is a doubly stochastic matrix, that is, for each row and each column, the sum of their entries equals one.

Random Sampling: The next main ingredient of our problem formulation is the description of the time instants at which the communication takes place between two sensor nodes connected by an edge. Corresponding to each edge $(i, j) \in \mathscr{E}$, it is stipulated that there is an increasing and divergent sequence $(\tau_k^{ij})_{k\in\mathbb{N}} \subset [0, +\infty[$ with $\tau_0^{ij} := 0$, and

- for each $(i, j) \in \mathscr{E}$, the sensor nodes $i, j \in \mathscr{V}$ transmit the value of their state estimate to each other at τ_k^{ij}, $k \in \mathbb{N}$.

In this article, we are interested in the case where the sampling times $(\tau_k^{ij})_{k\in\mathbb{N}}$ are generated *randomly*. Formally, we define

$$N_t^{ij} := \sup\big\{k \in \mathbb{N} \,\big|\, \tau_k^{ij} \leq t\big\} \quad \text{for } t \geq 0 \tag{6}$$

and assume in addition that, tor each $(i, j) \in \mathscr{E}$, $(N_t^{ij})_{t\geq 0}$ is a continuous-time stochastic process such that $\tau_{N_t^{ij}}^{ij} \to +\infty$ almost surely as $t \to +\infty$. The map $t \mapsto N_t^{ij}$ increments by 1 at random times, and it provides a description of the number of times the nodes i, j communicate with each other up to and including time t. For the sake of computational tractability, it is stipulated that

- For each $(i, j) \in \mathscr{E}$, $(N_t^{ij})_{t\geq 0}$ is an independent *Poisson process of intensity* $\lambda_{ij} > 0$. That is, $(N_t^{ij})_{t\geq 0}$ is a Markov process taking values in $\mathbb{N}$, has independent increments, and satisfies $N_0^{ij} = 0$, and for $h \searrow 0$ and $t \geq 0$,

$$\mathsf{P}\big(N_{t+h}^{ij} - N_t^{ij} = k \,\big|\, N_t^{ij}\big) = \begin{cases} 1 - \lambda_{ij}h + o(h) & \text{if } k = 0, \\ \lambda_{ij}h + o(h) & \text{if } k = 1, \\ o(h) & \text{if } k \geq 2, \end{cases}$$

  where the terms $o(h)$ do not depend on t.

Because of the arrival of new information at random times, the estimate $\widehat{x}_i$, $i \in \mathscr{V}$, gets updated. To describe this update rule, we associate with each node $i \in \mathscr{V}$, the process N_t^i,

$$N_t^i := \sum_{j\in\mathcal{N}_i} N_t^{ij}$$

so that N_t^i increments by one whenever node $i \in \mathscr{V}$ exchanges information with any of its neighbor. We recall

that N_t^i is also a Poisson process of intensity $\lambda_i := \sum_{j\in\mathcal{N}_i}\lambda_{ij}$. The times at which N_t^i gets incremented are denoted by $\tau_{N_t^i}$. We can now introduce the activation set $\mathcal{A}_t^i$,

$$\mathcal{A}_t^i := \left\{ j \in \mathcal{N}_i \,\middle|\, N_t^{ij} - N_{\underline{t}}^{ij} \neq 0, \underline{t} = \tau_{N_t^i - 1} \right\},$$

so that, at communication times $t_c = \tau_{N_t^i}$, the set $\mathcal{A}_{t_c}^i$ describes the neighbors of node $i \in \mathscr{V}$ that communicate their estimate to node $i \in \mathscr{V}$. Consequently, at $t_c = \tau_{N_t^i}$, we update the state estimate as follows:

$$\widehat{x}_i(t_c^+) = \sum_{j\in\mathcal{A}_{t_c}^i} \pi_{ij}\widehat{x}_j(t_c^-) + \Big(1 - \sum_{j\in\mathcal{A}_{t_c}^i}\pi_{ij}\Big)\widehat{x}_i(t_c^-), \quad (7)$$

where π_{ij} are the elements of the matrix Π introduced in (5). If $e_i := x - \widehat{x}_i$ denotes the estimation error, then because of this update rule, it is observed that,

$$e_i(t_c^+)e_i^\top(t_c^+) \leq \Big(1 - \sum_{j\in\mathcal{A}_{t_c}^i}\pi_{ij}\Big)e_i(t_c^-)e_i^\top(t_c^-) + \sum_{j\in\mathcal{A}_{t_c}^i}\pi_{ij}e_i(t_c^-)e_i^\top(t_c^-)$$

which is a direct consequence of the following lemma, whose proof appears in (Tanwani, 2022, Lemma III.3):

Lemma 1. Let m be a positive integer, and let $x_1, \ldots, x_m \in \mathbb{R}^n$. If $z := \sum_{j=1}^m \gamma_j x_j$ for some $\gamma_j \in [0,1]$, then

$$zz^\top \leq \sum_{j=1}^m \gamma_j x_j x_j^\top. \quad (8)$$

2.3 Summary of Filtering Algorithm

So far, we have specified the information available to each sensor node and a filtering algorithm, (4), (7), which uses this available information. If $\mathcal{Y}_t^i$ denotes the information available to sensor node $i \in \mathscr{V}$ up till time $t \in [0, +\infty[$, then we can write $\mathcal{Y}_t^i = \{(\mathrm{d}y_i(s), \widehat{x}_j(\tau_{N_s^i})) \,|\, s \leq t, j \in \mathcal{N}_i\}$. Our goal is to quantify the performance of these distributed filters by computing a bound on expected value of the error covariance matrices, that is, $\mathsf{E}[\mathsf{E}[(x(t) - \widehat{x}_i(t))(x(t) - \widehat{x}_i(t))^\top \,|\, \mathcal{Y}_t^i]]$, for $t \geq 0$. Here, the outer expectation is with respect to the random update times, and the inner expectation is with respect to the noise process in the state and output equation. The estimate computed by each node $i \in \mathscr{V}$ is obtained by executing the following steps:

- Integrate (4a) and (4b) over the interval $[\tau_{N_t^i}, \tau_{N_t^i+1}[$,
- At $t_c = \tau_{N_t^i}$, reset the state $\widehat{x}_i$ via (7), and set

$$P_i(t_c^+) = \big(1 - \sum_{j\in\mathcal{A}_{t_c}^i}\pi_{ij}\big)P_i(t_c^-) + \sum_{j\in\mathcal{A}_{t_c}^i}\pi_{ij}P_j(t_c^-) \quad (9)$$

with $P_i(0) \geq \mathsf{E}[(x(0) - \widehat{x}_i(0))(x(0) - \widehat{x}_i(0))]$.

As a first step in obtaining the desired bounds, one immediately observes that, for each $i \in \mathscr{V}$, if we fix the times at which node $i \in \mathscr{V}$ communicates with its neighbors, then

$$\mathsf{E}[(x(t) - \widehat{x}_i(t))(x(t) - \widehat{x}_i(t))^\top \,|\, \mathcal{Y}_t^i] \leq P_i(t), \quad t \geq 0 \quad (10)$$

where P_i is described by (4b), (9). The remaining task therefore is to compute expectation with respect to the distributions assigned to the times at which information between sensor nodes takes place.

3. MAIN RESULTS

The basic problem studied in this paper is the performance of the distributed filters proposed in the previous section. In particular, we want to relate the mean sampling rates λ_{ij}, corresponding to the edges $(i,j) \in \mathscr{E}$, with the bounds on the error covariance. As our first main result, we compute an upper bound on the expectation (with respect to sampling process N_t^i) of the error covariance matrices $\mathsf{E}[\mathsf{E}[(x(t) - \widehat{x}_i(t))(x(t) - \widehat{x}_i(t))^\top \,|\, \mathcal{Y}_t^i]]$, for $t \geq 0$.

Theorem 2. Consider system (1) with distributed measurements (3) and the corresponding hybrid filters (4), (7), (9) linked together by an undirected and connected graph $\mathcal{G} = (\mathscr{V}, \mathscr{E})$. For an edge $(i,j) \in \mathscr{E}$, if the communication between nodes $i, j \in \mathscr{V}$ takes place at random times generated by a Poisson process of intensity $\lambda_{ij} > 0$, then for each $i = 1, \cdots, N$, it holds that

$$\mathsf{E}[\mathsf{E}[(x(t) - \widehat{x}_i(t))(x(t) - \widehat{x}_i(t))^\top \,|\, \mathcal{Y}_t^i]] \leq \mathcal{P}_i(t), \quad (11)$$

where the matrix-valued function $\mathcal{P}_i : [0, \infty[\to \mathbb{R}^{n\times n}$ satisfies the differential equation

$$\dot{\mathcal{P}}_i = A\mathcal{P}_i + \mathcal{P}_i A^\top - \mathcal{P}_i H_i^\top V_i^{-1} H_i \mathcal{P}_i + BB^\top + \sum_{j\in\mathcal{N}_i}\lambda_{ij}\pi_{ij}(\mathcal{P}_j - \mathcal{P}_i). \quad (12)$$

The proof of Theorem 2 follows from the proof of (Tanwani, 2022, Theorem III.1, item 1)). The result of Theorem 2 provides a bound on the evolution of error covariance for each node in terms of a differential equation. These equations are quadratic (and hence nonlinear), driven by a constant term which corresponds to the noise level in the system, and are interconnected by some coupling term. Such systems in the literature are studied under the framework of heterogenous multi-agent systems since the dynamics of $\mathcal{P}_i$ are different for each $i \in \mathscr{V}$. In contrast to homogenous agents, consensus in heterogenous agents is not possible in general. However, one can get the states of all the agents close to desired accuracy by increasing the coupling strength. The next result relates to the asymptotic behavior of the coupled differential equations (12).

Theorem 3. For $i = 1, \ldots, N$, consider the matrix-valued equations (12) and assume that $\lambda_{ij} = \lambda$ is the same for each $(i,j) \in \mathscr{E}$. Let $S \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite matrix satisfying

$$0 = AS + SA^\top - \frac{1}{N}S\left(\sum_{i=1}^N H_i^\top V_i^{-1} H_i\right)S + BB^\top. \quad (13)$$

Then for every $\delta > 0$, there exists $\lambda > 0$ sufficiently large, such that the corresponding solution of (12) satisfies [1]

$$\limsup_{t\to\infty} \|\mathcal{P}_i(t) - S\| \leq \delta. \quad (14)$$

The proof of Theorem 3 is carried out in (Tanwani, 2021). To conclude this section, we provide some remarks about our main results.

Remark 4. The injection gains used in (4) over an interval $[\tau_{N_t^i}, \tau_{N_t^i+1}[$ do not need any information about how the other filters in the network choose their gains. Moreover, they minimize the value of P_i in (4b) over the class of linear time-varying gains. The latter statement follows from the fact that

[1] When taking the norm of a matrix, we refer to Frobenius norm.

$$(A-L_iH_i)P_i+P_i(A_i-L_iH_i)^\top+L_iV_iL_i^\top = (A-\overline{L}_iH_i)P_i + P_i(A_i-\overline{L}_iH_i)^\top + \overline{L}_iV_i\overline{L}_i^\top - (\overline{L}_i-L_i)V_i(\overline{L}_i-L_i)^\top \quad (15)$$

for any constant matrix $\overline{L}_i \in \mathbb{R}^{n\times p_i}$, and $L_i = P_iH_i^\top V_i^{-1}$. This shows that the filters (4) perform better than the constant linear gains proposed by the author in (Tanwani, 2022).

Remark 5. In Theorem 3, we basically study convergence of the differential equations (12) which contain quadratic nonlinearities. In general, such nonlinearities result in semiglobal convergence, that is, the solutions converge starting from initial conditions in a compact set. However, because of the minimum property described in Remark 4, we get global convergence with no restrictions on the initial condition. However, the convergence is not necessarily asymptotic, but only up to a neighborhood of a fixed point, which we often call practical convergence. The practical aspect of the convergence is unavoidable since we allow different noise covariance levels for each filter.

Remark 6. In the formulation of Theorem 2, since we associate a different Poisson process to each link, the nodes communicate with each other at different times. However, we associate the same sampling rate λ with each edge (i,j). The motivation for doing so is that, when we write the collective dynamics for each node, the last term in (12) is written as a scalar multiple of the Laplacian. If we assume that the process N_t^{ij} associated with edge (i,j) has intensity λ_{ij}, then the arguments required for establishing practical convergence are more involved and are not carried out in this paper. One would expect that if each $\lambda_{ij} > 0$ is large enough, then we do get practical convergence.

REFERENCES

Cao, M., Morse, A., and Anderson, B. (2008). Reaching a consensus in a dynamically changing environment: A graphical approach. *SIAM J. Control Optimization*, 47, 575–600.

DeGroot, M. (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69(345), 118–121.

Hespanha, J.P. (2014). Modeling and analysis of networked control systems using stochastic hybrid systems. *Annual Reviews in Control*, 38(2), 155 – 170.

Jadbabaie, A., Lin, J., and Morse, A. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6), 988–1001.

Moreau, L. (2005). Stability of multiagent systems with time-dependent communication links. *IEEE Trans. Automat. Control*, 50, 169–182.

Olfati-Saber, R. (2007). Distributed Kalman filtering for sensor networks. In *Proc. of 46th IEEE Conf. on Decision and Control*.

Olfati-Saber, R. (2009). Kalman-Consensus Filter : Optimality, stability, and performance. In *Proc. of 48th IEEE Conf. on Decision and Control*.

Tanwani, A. (2021). Error bounds for locally optimal distributed filters with random communication graphs. In *Proc. 60th IEEE Conf. Decision & Control*, 1609–1614.

Tanwani, A. (2022). Suboptimal filtering over sensor networks with random communication. *IEEE Transactions on Automatic Control*. DOI: 10.1109/TAC.2021.3116180.

Tanwani, A., Chatterjee, D., and Grüne, L. (2019). Performance bounds for stochastic receding horizon control with randomly sampled measurements. In *Proc. 58th IEEE Conf. Decision & Control*, 2330–2335.

Tanwani, A., Chatterjee, D., and Liberzon, D. (2018). Stabilization of continuous-time deterministic systems under random sampling: Overview and recent developments. In T. Başar (ed.), *Uncertainty in Complex Networked Systems*, 209–246. Springer Nature.

Tanwani, A. and Yufereva, O. (2020). Error covariance bounds for suboptimal filters with Lipschitzian drift and Poisson-sampled measurements. *Automatica*, 122.

Tsitsiklis, J., Bertsekas, D., and Athans, M. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31, 803–812.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Abstract: Positivity is Undecidable in Tensor Products of Free Algebras ⋆

Arthur Mehta * **William Slofstra** ** **Yuming Zhao** ***

* *University of Ottawa (amehta2@uottawa.ca)*
** *Institute for Quantum Computing and Department of Pure Mathematics, University of Waterloo (william.slofstra@uwaterloo.ca)*
*** *Institute for Quantum Computing and Department of Pure Mathematics, University of Waterloo (yuming.zhao@uwaterloo.ca)*

Abstract: This is an extended abstract of a paper "Positivity is undecidable in tensor products of free algebras" which is currently in preparation. In quantum information, we are interested in tensor products of free algebras and related algebras, since these tensor products model spatially separated subsystems with entanglement. The recent MIP*=RE result shows that it is undecidable to determine whether an element in the tensor product of two free group algebras is positive in all finite-dimensional representations. This shows that this tensor product is not RFD, resolving the Connes embedding problem. In this work, we show that these tensor products are also not archimedean-closed, by showing that it is undecidable to determine if an element of the tensor product is positive. The result also holds for tensor products of related algebras, like algebra of *-polynomials or the group algebra of a free product of abelian groups.

1. INTRODUCTION

An element of a *-algebra is said to be positive if it is positive in all *-representations. It is a natural problem to determine whether or not a given element of a *-algebra is positive, and if it is, find some certificate of positivity, like a sum of squares decomposition. For commutative algebras, this problem has a long history. For the group algebra of the free group, this problem can be solved in two ways:

(1) by Bakonyi and Timotin (2007), the algebra is archimedean-closed, so every positive element has a sum of squares decomposition, and
(2) by Choi (1980) the algebra is RFD, meaning that an element is positive if and only if it is positive in all finite-dimensional *-representations.

Thus for this algebra we can find certificates of both positivity and non-positivity. A similar solution is possible for the the algebra of *-polynomials and the algebra of contractions by Helton (2002) and Helton and McCullough (2004). Similarly, the semi-pre-C^*-algebra $\mathcal{A}(n,m)$ generated by positive elements p_a^x, $1 \leq x \leq n$, $1 \leq a \leq m$ and satisfying relations $\sum_{a=1}^m p_a^x = 1$ for all $1 \leq x \leq n$ is archimedean-closed and RFD by, e.g., Helton et al. (2012). (See also Ozawa (2013) for background on semi-pre-C^*-algebras.)

In quantum information, $\mathcal{A}(n,m)$ models a physical system with n possible measurements, each with m outcomes. The tensor product $\mathcal{A}(n,m) \otimes \mathcal{A}(n,m)$ models two spatially separated (but possibly entangled) subsystems of this form. The recent MIP*=RE result of Ji, Natarajan, Vidick, Wright, and Yuen shows that it is undecidable to determine whether an element in $\mathcal{A}(n,m) \otimes \mathcal{A}(n,m)$ is positive in all finite-dimensional representations (Ji et al. (2020)). This shows that this tensor product is not RFD, resolving the Connes embedding problem (see Ozawa (2013) for more background on this problem). In this work, we show that:

Theorem 1. There is a mapping from Turing machines M to elements $\alpha_M \in \mathcal{A}(n,m) \otimes \mathcal{A}(n,m)$ such that M halts if and only if α_M is not positive.

In other words, determining whether or not an element is positive in $\mathcal{A}(n,m) \otimes \mathcal{A}(n,m)$ is coRE-hard. As a corollary, this shows that $\mathcal{A}(n,m) \otimes \mathcal{A}(n,m)$ is not archimedean-closed. The result also holds for tensor products of related algebras, like algebra of *-polynomials, the algebra of contractions, or the group algebra of a free group.

⋆ WS acknowledges support from NSERC DG 2018-03968 and the Alfred P. Sloan Research Fellowship program.

REFERENCES

Bakonyi, M. and Timotin, D. (2007). Extensions of positive definite functions on free groups. *Journal of Functional Analysis*, 246, 31–49.

Choi, M.D. (1980). The full c*-algebra of the free group on two generators. *Pacific Journal of Mathematics*, 87(1), 41–48.

Helton, J.W. (2002). "Positive" Noncommutative Polynomials Are Sums of Squares. *Annals of Mathematics*, 156(2), 675–694.

Helton, J.W. and McCullough, S.A. (2004). A Positivstellensatiz for Non-Commutative Polynomials. *Trans. AMS*, 359(9), 3721–3737.

Helton, J., Klep, I., and McCullough, S. (2012). The convex positivstellensatz in a free algebra. *Advances in Mathematics*, 231(1), 516–534.

Ji, Z., Natarajan, A., Vidick, T., Wright, J., and Yuen, H. (2020). MIP*=RE. *unpublished.* ArXiv:2001.04383.

Ozawa, N. (2013). About the Connes embedding conjecture. *Japanese Journal of Mathematics*, 8, 147–183.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Robust Output-Feedback Adaptive Control of Finite Sets of Linear Systems ⋆

Olle Kjellqvist *

* *Lund University, Sweden, (e-mail: olle.kjellqvist@ control.lth.se).*

Abstract: This paper concerns the problem of bounded ℓ_2-gain adaptive control with noisy measurements for linear time-invariant systems with uncertain parameters belonging to a finite set. We show that it is necessary and sufficient to consider observer-based control with a multiple-observer structure consisting of one $\mathcal{H}_\infty$-observer paired with model fitness metric per candidate model for minimax optimality.

Keywords: adaptive control, machine learning and control

1. INTRODUCTION

The great control engineer is lazy; her models are simplified and imperfect, the operating environment may be poorly controlled — yet her solutions perform well. Robust control provides excellent tools to guarantee performance if the uncertainty is small Zhou and Doyle (1998). If the uncertainty is large, one can perform laborious system identification offline to reduce model uncertainty and synthesize a robust controller. An appealing alternative is to trade the engineering effort for a more sophisticated controller, particularly a learning-based component that improves controller performance as more data is collected. However, for such a controller to be implemented, it had better be robust to any prevalent unmodeled dynamics. Currently, there is considerable research interest in the boundary between machine learning, system identification, and adaptive control. For a review, see for example Matni et al. (2019). Most of the studies concern stochastic uncertainty and disturbances and assume perfect state measurements. Recently, works connecting to worst-case disturbances have started to appear. For example, non-stochastic control was introduced for known systems with unknown cost functions in Agarwal et al. (2019) and extended to unknown dynamics and output feedback, under the assumption of bounded disturbances and prior knowledge of a stabilizing proportional feedback controller in Simchowitz (2020). In Dean et al. (2019) the authors leverage novel robustness results to ensure constraint satisfaction while actively exploring the system dynamics. In this contribution, the focus is on worst-case models for disturbances and uncertain parameters as discussed in Didinsky and Basar (1994) and Vinnicombe (2004) and more recently in Rantzer (2021), but differ in that we consider output-feedback. See Figure 1 for an illustration of the considered problem. This paper extends Kjellqvist and Rantzer (submitted) to multiple input, multiple output systems and Theorem 1 in Rantzer (2021) to the output-feedback setting.

⋆ This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 834142 (ScalableControl).

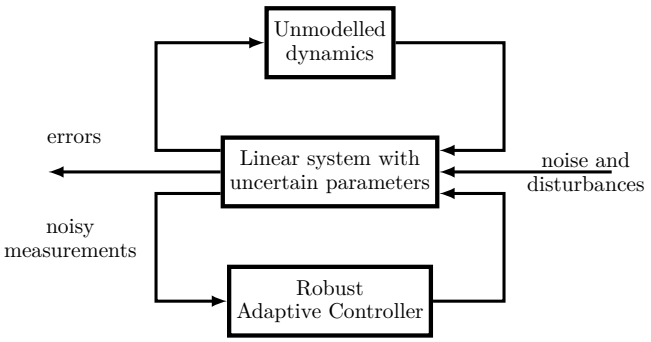


Fig. 1. For a finite set of linear time-invariant models, the Robust Adaptive Controller minimizes the ℓ_2-gain from noise and disturbances to errors for any realization of the unknown model parameters. This gain bound guarantees robustness to unmodelled dynamics.

The outline is as follows. We establish the notation in section 2. Section 3 defines the problem of finite-gain adaptive control and constructs a corresponding two-player dynamic game whose solution also solves the finite-gain adaptive control problem. We introduce the multi-observer as an information state in Section 4. Section 5 exploits the results from the previous setting to construct an equivalent full-information game and shows how it can be solved using dynamic programming. Concluding remarks are given in section 6.

2. NOTATION

The set of $n \times m$ matrices with real coefficients is denoted $\mathbb{R}^{n\times m}$. The transpose of a matrix A is denoted $A^\top$. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and a vector $x \in \mathbb{R}^n$ we use the expression $|x|_A^2$ as shorthand for $x^\top Ax$. We write $A \prec (\preceq)\ 0$ to say that A is positive (semi)definite. We refer to the value of a signal w at time t as $w(t)$ and use the shorthand notation $w_{0:t}$ for the sequence $(w(\tau))_{\tau=0}^t$.

3. PROBLEM FORMULATION

We consider uncertain linear systems of the form

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Gw(t), \quad x(0) = x_0 \\ y(t) &= Cx(t) + Hv(t), \quad t \geq 0, \\ M &= (A, B, C, G, H) \in \mathcal{M}, \end{aligned} \tag{1}$$

where the control signal $u(t) \in \mathbb{R}^{n_u}$ is generated by a causal output-feedback control policy

$$u(t) = \mu_t(y(0), \ldots, y(t), u(0), \ldots, u(t-1). \tag{2}$$

In (1), $x(t) \in \mathbb{R}^{n_x}$ is the state, $y(t) \in \mathbb{R}^{n_y}$ is the measurement, the model M is unknown but belongs to $\mathcal{M}$. For every model M, the matrix $H \in \mathbb{R}^{n_y \times n_v}$ is assumed to be surjective. Similarly to $\mathcal{H}_\infty$-control, we do not make explicit assumptions of the probability distributions of the disturbances $v(t) \in \mathbb{R}^{n_v}$ and $w(t) \in \mathbb{R}^{n_w}$. We are interested in control that makes the closed-loop system finite gain, with gain from (w, v, x_0) to $(Q^{1/2}x, R^{1/2}u)$ bounded above by γ. That is, the quantity

$$\begin{aligned} \alpha(T, \hat{x}_0) &:= |x(T)|_Q^2 + \sum_{t=0}^{T-1} \left(|x(t)|_Q^2 + |u(t)|_R^2 \right) \\ &- \gamma^2 \sum_{t=0}^{T-1} \left(w(t)^2 + v(t)^2 \right) - \gamma^2 |v(T)|^2 - |x(0) - \hat{x}_0^2|_{P_M}^2 \end{aligned} \tag{3}$$

must be bounded for all $T \geq 0$, any admissible disturbances, initial state and the possible realizations M of (1). P_M quantifies prior information on the initial state and is taken as a positive solution to the Riccati equation

$$\begin{aligned} P_M &= \left(AX_M^{-1}A^\top + \gamma^{-2} GG^\top \right)^{-1}, \\ X_M &= P_M + \gamma^2 C^\top (HH^\top)^{-1} C - Q). \end{aligned} \tag{4}$$

We cannot evaluate condition (3) directly since disturbances w, noise v and the active plant M are unknown. We can, however, evaluate the worst case (the supremum), and will consider the dynamic game with dynamics as in (1), (2) and the objective:

$$J_\star(\hat{x}_0) = \inf_\mu \sup_{w_{0:T-1}, v_{0:T}, x_0, M, T} \alpha(T, \hat{x}_0). \tag{5}$$

This problem setup is similar to a standard linear-quadratic game Didinsky and Basar (1994) but differs because the adversary selects the active model $M \in \mathcal{M}$.

4. A MULTI-OBSERVER INFORMATION STATE

Consider the case when the data, $(y_{0:T}, u_{0:T-1})$, is generated by the dynamics (1) and (2) under model M and control policy μ. Let $\alpha_M(T, \hat{x}_0)$ be the largest value of $\alpha(T, \hat{x}_0)$ consistent with the data and the dynamics,

$$\begin{aligned} \alpha_M(T) = \sup_{w_{0:T-1}, v_{0:T}, x_0} \Big\{ &\alpha(T) : \\ &(y_{0:T}, u_{0:T-1}) \text{ generated by (1) and (2) under } M \Big\}. \end{aligned} \tag{6}$$

We can partition the supremum in (5) into

$$J_\star(\hat{x}_0) = \inf_\mu \sup_{y_{0:T}, M, T} \alpha_M(T, \hat{x}_0).$$

Define the state-dependent past cost by

$$\begin{aligned} W_M(t+1, x, \hat{x}_0) = \sup_{w_{0:t}, v_{0:t}, x_0} \Bigg\{ &\sum_{\tau=0}^{t} \left(|x(\tau)|_Q^2 + |u(\tau)_R^2 \right) \\ &- \gamma^2 \sum_{\tau=0}^{t} \left(|w(\tau)|^2 + |v(\tau)^2 \right) - |x(0) - \hat{x}_0|_{P_M}^2 \Bigg\}, \end{aligned} \tag{7}$$

where the supremum is taken with respect to the dynamics (1) and a fixed trajectory $(y_{0:t}, u_{0:t})$. Then α_M becomes

$$\begin{aligned} &\alpha_M(T, \hat{x}_0) \\ &= \sup_x \left\{ |x|_Q^2 - \gamma^2 |Cx - y(T)|_{(HH^\top)^{-1}}^2 + W_M(T, x, \hat{x}_0) \right\}. \end{aligned} \tag{8}$$

In (Basar and Bernhard, 1995, Chapter 6), the authors show how to express W_M recursively. We summarize these results in the following Lemma:

Lemma 1. Given a known model $M \in \mathcal{M}$, a positive quantity γ, positive definite matrices $Q \in \mathbb{R}^{n_x \times n_x}$ and $R \in \mathbb{R}^{n_u \times n_u}$. Assume that the Riccati equation (4) has a positive definite solution P_M such that X_M is positive definite. For a fixed trajectory $(u_{0:T}, y_{0:T})$, W_M in (7) obeys

$$W_M(t, x, \hat{x}_0) = -|x - \hat{x}_M(t)|_{P_M}^2 + l_M(t). \tag{9}$$

The observer state $\hat{x}_M(t)$ is the solution to the dynamical system

$$\begin{aligned} \hat{x}_M(t+1) &= A\hat{x}_M(t) + Bu(t) \\ &\quad + K_M(y(t) - C\hat{x}_M(t)) + \hat{w}_M(t) \\ K_M &= \gamma^2 A X_M^{-1} C^\top (HH^\top)^{-1}, \\ \hat{w}_M(t) &= A X_M^{-1} Q \hat{x}_M(t), \end{aligned} \tag{10}$$

and l_M obeys the recurrence relation

$$\begin{aligned} l_M(t+1) &= l_M(t) - |\hat{x}_M(t)|_{P_M}^2 - \gamma^2 |y(t)|_{(HH^\top)^{-1}}^2 \\ &+ |u(t)|_R^2 + |P_M \hat{x}_M(t) + \gamma^2 C^\top (HH^\top)^{-1} y(t)|_{X_M^{-1}}^2. \end{aligned} \tag{11}$$

The initial conditions are $(\hat{x}_M(0), l_M(0)) = (\hat{x}_0, 0)$.

We can evaluate the supremum in (8),

$$\begin{aligned} \alpha_M(T) &= l_M(T) - |\hat{x}_M(T)|_{P_M}^2 - \gamma^2 |y(T)|_{(HH^\top)^{-1}}^2 \\ &\quad + |P_M \hat{x}_M(T) + \gamma^2 C^\top (HH^\top)^{-1} y(T)|_{X_M^{-1}}^2 \end{aligned} \tag{12}$$

Thus, the collection of model performance quantities $\{l_M(t) : M \in \mathcal{M}\}$ is sufficient information to evaluate the finite-gain condition at time $t-1$ together with the observer states $\{\hat{x}_M(t) : M \in \mathcal{M}\}$, and together with $y(t)$ we have all the information necessary to compute $l_M(t+1)$. In other words, we have compressed the information in the sequences $(y_{0:t-1}, u_{0:t-1})$ of increasing length, to one quantity l_M and one state-vector $\hat{x}_M$ per model M.

5. OUTPUT-FEEDBACK MINIMAX DYNAMIC PROGRAMMING

In this section we will construct a full-information dynamic game, whose value equals (5). We then show that the game can be solved using dynamic programming in Theorem 2. Consider the game defined by the objective function

$$\inf_\eta \sup_{y, T, M} \left\{ |\hat{x}_M(T)|_{(Q^{-1} - P_M^{-1})^{-1}}^2 + l_M(T) \right\} \tag{13}$$

and the observer dynamics described in (10) and (11). The control signal is computed according to

$$u(t) = \eta(\mathcal{O}(t), y(t)),$$

where $\mathcal{O}$ is the multi-observer state,

$$\mathcal{O}(t) := \{(\hat{x}_M(t), l_M(t)) : M \in \mathcal{M}\}. \tag{14}$$

Define the Bellman operator $\mathcal{F}$ and the corresponding value iterations $V_0, V_1, \dots$ by

$$\mathcal{F}V(\mathcal{O}(t)) = \max_y \min_u V(\mathcal{O}(t+1)), \tag{15}$$

and

$$\begin{aligned} V_0(\mathcal{O}) &:= \max_{M\in\mathcal{M}} \left\{ |\hat{x}_M|^2_{(Q^{-1}-P_M^{-1})^{-1}} + l_M \right\} \\ V_{t+1}(\mathcal{O}) &= \mathcal{F}V_t(\mathcal{O}). \end{aligned} \tag{16}$$

where $\hat{x}_M(t+1)$ and $l_M(t+1)$ are computed as in (10) and (11).

The following theorem is an extension of (Rantzer, 2021, Theorem 1) to the output feedback setting. Both results rely on establishing an information state that replaces the uncertainty in the dynamics by a terminal cost. The theorem establishes that the values (5) and (13) are equal, and that the value iteration can be used to construct optimal and suboptimal controllers for (5). The corresponding controller architecture is illustrated in Fig. 2.

Theorem 2. Given a finite set $\mathcal{M}$ of linear dynamical systems, positive definite matrices $Q \in \mathbb{R}^{n_x \times n_x}$, $R \in \mathbb{R}^{n_u \times n_u}$ and a positive quantity γ. Assume that the Riccati equations (4), have positive definite solutions $P_M \succ Q$ for each model $M \in \mathcal{M}$. Then the values of (5) and (13) are finite if and only if the value iteration $V_0, V_1, \dots$ defined in (5) is bounded. If the value iteration is bounded, the limit $V_\star = \lim_{k\to\infty} V_k$ exists and the values of (5) and (13) are both equal to $V_\star(\mathcal{O}_0)$, where $\mathcal{O}_0 = \{(\hat{x}_0, 0) : M \in \mathcal{M}\}$ is the initial observer state as in (14).

Furthermore, denote by $\eta_\star$ the minimizing argument of $\mathcal{F}V_\star$, then $\eta_\star$ is optimal for (13), and the policy

$$\mu_\star(y(0), \dots, y(t), u(0), \dots, u(t-1)) := \eta_\star(\mathcal{O}(t), y(t)),$$

is optimal for (5).

If there exists a function $\bar{V} \geq V_0$, and a control policy $\bar{\eta}$ such that

$$\max_y \bar{V}(\mathcal{O}(t+1)) \leq \bar{V}(\mathcal{O}(t)),$$

where $u_t = \bar{\eta}(\mathcal{O}(t), y(t))$, then the values (5) and (13) are bounded above by $\bar{V}(\mathcal{O}_0)$. The control policy $u_t = \bar{\eta}(\mathcal{O}(t), y(t))$ achieves

$$J_{\bar{\mu}}(\hat{x}_0) := \sup_{w_{0:T-1}, v_{0:T}, x_0, M, T} \alpha(T, \hat{x}_0) \leq \bar{V}(\mathcal{O}_0).$$

The proof follows closely that of (Rantzer, 2021, Theorem 1), but differs in the dynamics of the information state.

Proof. Consider,

$$\begin{aligned} \mathcal{F}V_0(\mathcal{O}) &\geq \max_{M,y} \Big\{ l_M - |\hat{x}_M|^2_{P_M} \\ &\quad - \gamma^2 |y|^2_{(HH^\top)^{-1}} + |P_M \hat{x}_M + \gamma^2 C^\top (HH^\top)^{-1} y|^2_{X_M^{-1}} \Big\} \\ &= \max_M \{ |\hat{x}_M|^2_{(Q^{-1}-P_M^{-1})^{-1}} + l_M \} = V_0(\mathcal{O}). \end{aligned}$$

As $\mathcal{F}$ is monotone increasing in V, the sequence $V_0, V_1, V_2, \dots$ is monotonically non-decreasing. For any fix $T \geq 0$, $J_\star(\hat{x}_0)$ is bounded below by the finite-time objective

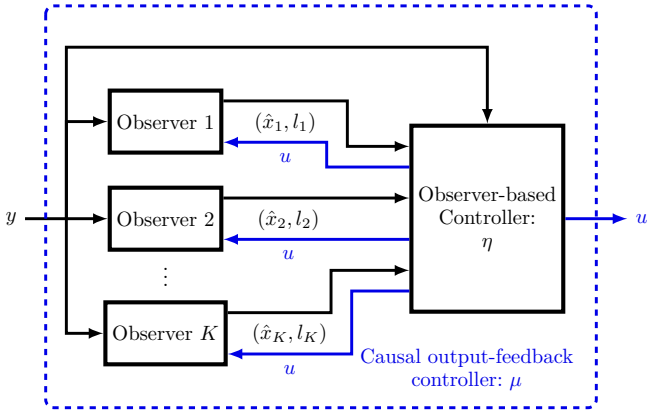


Fig. 2. Illustration of the proposed multi-observer control architecture. One $\mathcal{H}_\infty$-observer for each model $M \in \mathcal{M}$ outputs its state-estimate $\hat{x}_M(t)$ and model performance quantity $l_M(t)$ at each time t. The observer-based control policy η maps the observer states, performance quantities and the current output to the actuator signal $u(t)$.This control architecture forms the causal control policy $\mu(y_{0:t}, u_{0:(t-1)})$ that solves (5).

$$\inf_\mu \sup_{w_{0:T-1}, v_{0:T}, x_0, M} \alpha(T, \hat{x}_0). \tag{17}$$

Equation (17) is increasing in T, and $J_\star$ is obtained in the limit. By Lemma 1, (17) is equal to

$$\inf_\mu \sup_{y,M} \left\{ |\hat{x}_M(T)|^2_{(Q^{-1}-P_M^{-1})^{-1}} + l_M(T) \right\}. \tag{18}$$

Standard dynamic programming arguments show that the value of (18) is equal to $V_{T+1}(\mathcal{O}_0)$. This proves that (5) has a finite value if and only if $\{V_k(\mathcal{O}_0)\}_{k=0}^\infty$ is upper bounded, and that the value is equal to the limit $V_\star(\mathcal{O}_0) = \lim_{k\to\infty} V_k(\mathcal{O}_0)$.

To show that $V_\star(\mathcal{O})$ exists for any observer state $\mathcal{O}$ we first note that for any other $\hat{x}_0' \in \mathbb{R}^{n_x}$,

$$-\max_M |\hat{x}_0 - \hat{x}_0'|_{P_M} \leq J_\star(\hat{x}_0) - J_\star(\hat{x}_0') \leq \max_M |\hat{x}_0 - \hat{x}_0'|_{P_M}.$$

This shows that the value iteration remains bounded if we replace the initial states $\hat{x}_0$. To see that the iteration is bounded for arbitrary values of the performance quantities, let $\mathcal{O}' = \{(\hat{x}_0, l_M') : M \in \mathcal{M}\}$, then

$$\min_{M\in\mathcal{M}} l_M' \leq V_k(\mathcal{O}') - V_k(\mathcal{O}_0) \leq \max_{M\in\mathcal{M}} l_M'.$$

We conclude that if $V_k(\mathcal{O}_0)$ is bounded for $k = 1, 2, \dots$, so is $V_k(\mathcal{O}')$ and that $V_\star(\mathcal{O})$ exists for all multi-observer states $\mathcal{O}$.

If (13) is finite, then V_t is bounded above by (13), so the limit $V_\star$ is finite. Conversely, if $V_0 \leq \bar{V} \leq \infty$ and $\mathcal{F}_{\bar{\eta}}\bar{V} \leq \bar{V}$, we may define the sequence $\hat{V}_0, \hat{V}_1, \dots$ by the recursion $\hat{V}_0 := V_0$ and

$$\hat{V}_{t+1} := \mathcal{F}_{\bar{\eta}} \hat{V}_t.$$

By construction, (13) is bounded above by $\lim_{t\to\infty} \hat{V}_t(\mathcal{O}_0)$. By induction, $\bar{V} \geq \hat{V}_k$ for all k, so $V_\star \leq \lim_{k\to\infty} \leq \bar{V}$. This proves that $V_\star(\mathcal{O}_0) \leq J_{\bar{\mu}}(\hat{x}_0) \leq \bar{V}(\mathcal{O}_0)$. In particular, the control law is optimal if $\bar{V} = V_\star$.

6. CONCLUSIONS

We have shown that it is both necessary and sufficient to consider multi-observer-based feedback for minimax optimal control of a linear system with uncertain parameters belonging to a finite set. Such a controller guarantees that the ℓ_2-gain from disturbances to output is bounded. In future research, we aim to leverage these results together with approximate dynamic programming to synthesize control policies with a guaranteed upper bound on the gain from disturbances to error.

REFERENCES

Agarwal, N., Bullins, B., Hazan, E., Kakade, S., and Singh, K. (2019). Online control with adversarial disturbances. In K. Chaudhuri and R. Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 111–119. PMLR.

Basar, T. and Bernhard, P. (1995). *H_∞-Optimal Control and Related Minimax Design Problems — A dynamic Game Approach*. Birkhauser.

Dean, S., Tu, S., Matni, N., and Recht, B. (2019). Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, 5582–5588. doi:10.23919/ACC.2019.8814865.

Didinsky, G. and Basar, T. (1994). Minimax adaptive control of uncertain plants. In *Proceedings of 1994 33rd IEEE Conference on Decision and Control*, volume 3, 2839–2844 vol.3. doi:10.1109/CDC.1994.411368.

Kjellqvist, O. and Rantzer, A. (submitted). Learning-enabled robust control with noisy measurements. In *Proceedings of the 4th Conference on Learning for Dynamics and Control*, Proceedings of Machine Learning Research. PMLR. URL `https://github.com/kjellqvist/noisy-lerc`.

Matni, N., Proutiere, A., Rantzer, A., and Tu, S. (2019). From self-tuning regulators to reinforcement learning and back again. 3724–3740. doi:10.1109/CDC40024.2019.9029916.

Rantzer, A. (2021). Minimax adaptive control for a finite set of linear systems. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, 893–904. PMLR.

Simchowitz, M. (2020). Making non-stochastic control (almost) as easy as stochastic. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, 18318–18329. Curran Associates, Inc.

Vinnicombe, G. (2004). Examples and counterexamples in finite l2-gain adaptive control.

Zhou, K. and Doyle, J.C. (1998). *Essentials of Robust Control*. Prentice-Hall.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On the stabilizing properties of nonlinear MPC with arbitrary positive definite cost functions

Mircea Lazar *

* *Eindhoven University of Technology, Eindhoven, Netherlands*
(e-mail: m.lazar@tue.nl)

Abstract: Recently, a unifying approach to the stability analysis of nonlinear model predictive controllers (MPC) with arbitrary positive definite cost functions has been presented based on dissipativity theory. We have established that regardless of the choice of the positive definite cost function, the resulting value function always satisfies a dissipation inequality. This led to less conservative stability conditions for nonlinear MPC that do not require monotonic decrease of the optimal cost function along closed–loop trajectories. In this extended abstract we recall these results and we analyze recursive feasibility, which has not yet been addressed. To this end we use a control contractive terminal set and an adaptive prediction horizon, without adding a terminal cost.

Keywords: Predictive control, Recursive feasibility, Control contractive sets, Stability of nonlinear systems.

1. INTRODUCTION

Model predictive control (MPC) computes a control action online by minimizing a finite–horizon cost function subject to constraints. Standard stabilizing conditions for MPC (Mayne, 2014; Rawlings et al., 2017) utilize the corresponding optimal cost function as a monotonic Lyapunov function (LF). These conditions can be classified as follows (Mayne, 2014): *(i)* terminal state and input equal to steady–state values, e.g., see (Bemporad et al., 1994); *(ii)* local control Lyapunov function (CLF) as terminal cost and terminal state constrained to a controlled invariant set, e.g., see (Mayne, 2013); *(iii)* local CLF as terminal cost, no terminal state constraint and sufficiently long prediction horizon, e.g., see (Mayne, 2001; Limon et al., 2006); and *(iv)* no terminal cost, no terminal constraint, stage cost satisfies an asymptotic controllability assumption and sufficiently long prediction horizon, e.g., see (Grüne, 2012; Boccia et al., 2014). These approaches to stability analysis of nominal MPC consider a known, time–invariant steady–state equilibrium, typically chosen as the origin, and they require monotonically decreasing MPC cost functions.

However, predictive control with arbitrary positive definite cost functions and prediction horizon values often yields converging trajectories without imposing the above mentioned conditions, allowing even for non–monotonic cost functions. Also, industrial practice for MPC tuning typically employs a sufficiently long prediction horizon, yielding converging trajectories, instead of enforcing a monotonically decreasing cost. It is thus of interest to identify conditions under which predictive control with arbitrary positive definite cost functions results in stable closed–loop systems and recursively feasible optimization problems.

Recently, (Lazar, 2021) presented a unifying approach to the stability analysis of nonlinear model predictive control (MPC) with arbitrary positive definite cost functions based on dissipativity theory. Therein, it was established that regardless of the choice of the positive definite cost function, the resulting value function always satisfies a dissipation inequality. Then, it became clear that nonlinear MPC with arbitrary positive definite stage costs is stabilizing whenever the supply function induced by the stage cost is negative definite along closed–loop trajectories.

In this extended abstract we provide a summary of the main result of (Lazar, 2021) and we address the problem of guaranteeing recursive feasibility for nonlinear MPC with arbitrary positive definite cost functions. We show that recursive feasibility and asymptotic stability of nonlinear MPC with an arbitrary positive definite cost function can be achieved by using control contractive terminal sets, as proposed in (Limon et al., 2005). Therein the focus was on enlarging the region of attraction of stabilizing nonlinear MPC with a special terminal cost by using a sequence of controllable or contractive terminal sets. In this work we show that a control contractive terminal set suffices to guarantee recursive feasibility and asymptotic stability for nonlinear MPC with arbitrary positive definite cost functions, i.e., without using a special terminal cost function. This may require however dynamic, online adaptation of the prediction horizon.

2. PRELIMINARIES

Let $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{N}$ denote the field of real numbers, the set of non–negative reals and the set of natural numbers, respectively. For a vector $x \in \mathbb{R}^n$, $\|x\|$ denotes an arbitrary p-norm, $p \in \mathbb{N}_{\geq 1} \cup \infty$. A function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ *belongs to class* $\mathcal{K}$ if it is continuous, strictly increasing and $\varphi(0) = 0$.

A function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ *belongs to class* $\mathcal{K}_\infty$ if $\varphi \in \mathcal{K}$ and $\lim_{s\to\infty} \varphi(s) = \infty$.

Consider a discrete–time dynamical system

$$x(k+1) = f(x(k), u(k)), \quad k \in \mathbb{N}, \tag{1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a suitable function that is zero at zero. We assume that the origin is a stabilizable equilibrium for (1). The system variables are constrained to compact sets with the origin in their interior, i.e. $(x, u) \in \mathbb{X} \times \mathbb{U}$. We assume that $\mathbb{X}$ is a constrained control invariant set, i.e., for all $x \in \mathbb{X}$, there exists a $u := \kappa(x) \in \mathbb{U}$ with $\kappa(0) = 0$ such that $f(x, u) \in \mathbb{X}$. For brevity, we refer to (Lazar, 2006) for the definition of asymptotic Lyapunov stability in $\mathbb{X}$.

Consider the following discrete–time dissipation inequality

$$V(x(k+1)) - V(x(k)) \leq s(x(k), u(k)), \quad \forall k \in \mathbb{N}, \tag{2}$$

where $V : \mathbb{R}^n \to \mathbb{R}$ is a positive definite storage function and $s : \mathbb{R}^n \to \mathbb{R}$ is a supply function that is bounded on bounded sets and $s(0, 0) = 0$.

In what follows we relax the latter requirement and provide a less conservative condition on the supply function s for inferring asymptotic stability from dissipativity. To this end, the following assumptions are instrumental.

Assumption 1. Controlled $\mathcal{K}$–boundedness: For the systems dynamics $f(\cdot, \cdot)$ and the state–feedback control law $u(k) = \kappa(x(k))$ that renders $\mathbb{X}$ controlled invariant it holds that $\|f(x, \kappa(x))\| \leq \sigma(\|x\|)$ for all $x \in \mathbb{X}$ and some $\sigma \in \mathcal{K}$.

Assumption 2. The storage function V satisfies

$$\alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|), \quad \forall x \in \mathbb{X}, \tag{3}$$

for some $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$.

Theorem 3. (Lazar, 2021) Let Assumption 1 and Assumption 2 hold and let $\alpha_s \in \mathcal{K}_\infty$. Suppose that $\{u(k)\}_{k\in\mathbb{N}} = \{\kappa(x(k))\}_{k\in\mathbb{N}}$ is such that the dissipation inequality (2) holds for all $x(0) \in \mathbb{X}$ and all $k \in \mathbb{N}$ and $(x(k), u(k)) \in \mathbb{X} \times \mathbb{U}$ for all $k \in \mathbb{N}$. Furthermore, suppose that there exists a $M \in \mathbb{N}_{\geq 1}$ such that for all $x(0) \in \mathbb{X}$ it holds that

$$\sum_{i=0}^{M-1} s(x(k+i), u(k+i)) \leq -\alpha_s(\|x(k)\|), \quad \forall k \in \mathbb{N}. \tag{4}$$

Then the origin of system (1) in closed–loop with $u(k) = \kappa(x(k))$ is asymptotically Lyapunov stable in $\mathbb{X}$.

3. STABILITY OF NONLINEAR MPC

In this section we analyze stability of state–space nonlinear MPC via Theorem 3. We will use the notation $x(i|k)$ to denote the predicted state at time $i \in \mathbb{N}_{[1,N]}$, given measured state $x(0|k) := x(k)$, and similarly $u(i|k)$ to denote the predicted input at time $i \in \mathbb{N}_{[0,N-1]}$, while setting $u(k) := u(0|k)$. The prediction model is a copy of (1), i.e.,

$$x(i+1|k) = f(x(i|k), u(i|k)), \quad i \in \mathbb{N}_{[0,N-1]},\ k \in \mathbb{N}.$$

Defining $\mathbf{u}(k) := \{u(0|k), \ldots, u(N-1|k)\}$ yields the MPC optimization problem:

$$\min_{\mathbf{u}(k)} J(x(k), \mathbf{u}(k)) := l_N(x(N|k)) + \sum_{i=0}^{N-1} l(x(i|k), u(i|k)) \tag{5a}$$

subject to constraints:

$$x(i+1|k) = f(x(i|k), u(i|k)), \quad \forall i \in \mathbb{N}_{[0,N-1]}, \tag{5b}$$

$$(x(i+1|k), u(i|k)) \in \mathbb{X} \times \mathbb{U}, \quad \forall i \in \mathbb{N}_{[0,N-1]}. \tag{5c}$$

In equation (5a) $l_N : \mathbb{R}^n \to \mathbb{R}_+$ denotes the terminal cost and $l : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$ denotes the stage cost. The following assumptions are required to ensure Assumption 2 for the nonlinear MPC optimal cost.

Assumption 4. There exists $\alpha_{1,l} \in \mathcal{K}_\infty$ such that

$$\alpha_{1,l}(\|x\|) \leq l(x, u), \quad \forall (x, u) \in \mathbb{X} \times \mathbb{U}.$$

Let $\mathbf{u}^*(k)$ and $\mathbf{x}^*(k) := \{x^*(i|k)\}_{i\in\mathbb{N}_{[1,N]}}$ denote optimal input and state trajectories obtained by solving the MPC problem (5) at time k. Let $\mathbf{u}^s(k) := \{u^*(1|k), \ldots, u^*(N-1|k), u^s(N|k)\}$ with $u^s(N|k) \in \mathbb{U}$ and $x^s(N+1|k) := f(x^*(N|k), u^s(N|k)) \in \mathbb{X}$ denote a shifted input trajectory constructed at time k from the optimal trajectory computed at time k. Let $J(x(k), \mathbf{u}^*(k))$ denote the optimal cost function or value function at time k.

Assumption 5. There exists $\alpha_{2,J} \in \mathcal{K}_\infty$ such that

$$J(x, \mathbf{u}^*) \leq \alpha_{2,J}(\|x\|), \quad \forall x \in \mathbb{X}.$$

Next, we state the nominal MPC dissipativity result.

Theorem 6. (Lazar, 2021) Suppose that $\mathbb{X} \times \mathbb{U}$ is constrained control invariant for system (1). Let the stage cost be any positive definite function that satisfies Assumption 4. For the terminal cost either of the following choices can be made: any positive definite function that is zero at zero or a zero terminal cost. Define the storage function $V(x(k)) := J(x(k), \mathbf{u}^*(k))$ and the supply function

$$\begin{aligned} s(x(k), u(k)) :=& l_N(x^s(N+1|k)) + l(x^*(N|k), u^s(N|k)) \\ &- l_N(x^*(N|k)) - l(x(k), u(k)), \quad k \in \mathbb{N}. \end{aligned} \tag{6}$$

Then for all $x(0) \in \mathbb{X}$, the dissipation inequality

$$V(x(k+1)) - V(x(k)) \leq s(x(k), u(k)), \quad \forall k \in \mathbb{N} \tag{7}$$

holds along the trajectories of system (1) in closed–loop with $u(k) := u^*(0|k)$ obtained by solving problem (5).

Let the MPC closed–loop system denote system (1) in closed–loop with $u^*(0|k)$ obtained by solving (5). Next we state the nominal MPC asymptotic stability result.

Corollary 7. (Lazar, 2021) Let the stage cost in the MPC problem (5) satisfy Assumption 4, suppose Assumption 5 holds and assume $\mathbb{X} \times \mathbb{U}$ is constrained control invariant. Suppose that there exists a $\rho \in \mathcal{K}_\infty$ and an $M \in \mathbb{N}_{\geq 1}$ such that Assumption 1 holds for the trajectories of the MPC closed–loop system, and the MPC supply function defined in (6) satisfies for all $x(0) \in \mathbb{X}$

$$\sum_{i=0}^{M-1} s(x(k+i), u(k+i)) \leq -\rho \circ l(x(k), u(k)), \quad \forall k \in \mathbb{N}. \tag{8}$$

Then the origin of the MPC closed–loop system is asymptotically Lyapunov stable in $\mathbb{X}$.

Relation with existing stabilizing conditions: The stabilizing conditions corresponding to approach *(i)* (terminal

equality constraint) force $x(N|k) = 0$, which allows setting $u^s(N|k) = 0$ and $x^s(N+1|k) = 0$. Hence, $s(x(k), u(k)) = -l(x(k), u(k))$ and condition (8) holds with $M = 1$ and $\rho = \mathrm{id}$. Approach *(ii)* (terminal set constraint) constructs $u^s(N|k) = h(x^*(N|k))$ such that $l_N(x)$ is a Lyapunov function for $x(k+1) = f(x(k), h(x(k)))$ which yields

$$l_N(x^s(N+1|k)) - l_N(x^*(N|k)) + l(x^*(N|k), u^s(N|k)) \leq 0.$$

Hence, $s(x(k), u(k)) \leq -l(x(k), u(k))$ and condition (8) holds with $M = 1$ and $\rho = \mathrm{id}$. Approach *(iii)* (MPC without terminal constraint and with CLF terminal cost) requires that $l_N(x)$ is a control Lyapunov function, i.e.,

$$\min_{u^s(N|k)} l_N(x^s(N+1|k)) - l_N(x^*(N|k)) + l(x^*(N|k), u^s(N|k))$$

should be less than or equal to zero. This implies existence of $u^s(N|k)$ such that $s(x(k), u(k)) \leq -l(x(k), u(k))$ and condition (8) holds with $M = 1$ and $\rho = \mathrm{id}$. Approach *(iv)* (MPC without terminal constraint and with zero terminal cost) uses the property (see (Grüne, 2012, Proposition 3.4))

$$V(x(k+1)) - V(x(k)) \leq -\alpha l(x(k), u(k)),$$

where $\alpha \in (0, 1]$ is a constant. Taking a zero terminal cost and requiring a monotonic decrease of V, as in (Grüne, 2012, Proposition 3.4), corresponds to existence of $u^s(N|k)$ such that

$$\begin{aligned} s(x(k), u(k)) &= l(x^*(N|k), u^s(N|k)) - l(x(k), u(k)) \\ &\leq -\rho \circ l(x(k), u(k)), \end{aligned}$$

which implies that (8) holds with $M = 1$ and $\rho < \mathrm{id}$.

Next, we briefly indicate how to construct a shifted input trajectory that verifies the developed stabilizing conditions. To begin with, we opt for using a terminal cost equal to the stage cost for the state, i.e. assume that $l(x, u) = l(x, 0) + l(0, u)$ and set $l_N(x) := l(x, 0)$. Then the MPC supply function becomes:

$$\begin{aligned} s(x(k), u(k)) :=& l(x^s(N+1|k), 0) + l(x^*(N|k), u^s(N|k)) \\ &- l(x^*(N|k), 0) - l(x(k), u(k)) \\ &= l(x^s(N+1|k), u^s(N|k)) - l(x(k), u(k)). \end{aligned} \tag{9}$$

At every time $k \in \mathbb{N}$, given $x(k)$, after $\mathbf{u}^*(k)$ was calculated, we need to check that there exists a $u^s(N|k) \in \mathbb{U}$ such that

$$\begin{aligned} &s(x(k), u(k)) = \\ &l(f(x^*(N|k), u^s(N|k)), u^s(N|k)) - l(x(k), u^*(0|k)) \\ &\leq -\rho \circ l(x(k), u^*(0|k)), \end{aligned} \tag{10}$$

for some $\rho \in \mathcal{K}_\infty$ with $\rho < \mathrm{id}$. If the above condition holds, then the cyclically neutral supply condition also holds.

The above inequality can be further relaxed by exploiting the accumulated supply over the previous M time steps, where the current time must satisfy $k \geq M + 1$, i.e.,

$$\begin{aligned} &\sum_{i=-M+1}^{0} s(x(k+i), u(k+i)) \\ &\leq -\rho \circ l(x(k-M+1), u^*(0|k-M+1)). \end{aligned} \tag{11}$$

Indeed, if we let $M = 1$ in the above inequality, we recover (10).

4. RECURSIVE FEASIBILITY ANALYSIS

The following results are a new addition to (Lazar, 2021). In order to guarantee recursive feasibility, consider the following terminal set characterization for system (1):

$$\mathbb{X}_T := \{x \in \mathbb{X} \ : \ g^q(\mathcal{X}, x) \leq c\},$$

where $c > 0$ is such that $\mathbb{X}_T \subseteq \mathbb{X}$, $\mathcal{X}$ is a proper C–set with gauge function $g(\cdot, \cdot)$ (Blanchini et al., 2015), and $q = 1$ or 2 for ellipsoidal proper C-sets.

Assumption 8. There exists an admissible control law $\kappa : \mathbb{X}_T \to \mathbb{U}$ continuous at the origin and zero at zero such that the set $\mathbb{X}_T$ is control λ–contractive with $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$, $\lambda(0) = 0$ and $\lambda < \mathrm{id}$, for the dynamics (1), i.e., for all $x \in \mathbb{X}_T$ it holds that $g^q(\mathcal{X}, f(x, \kappa(x)) \leq \lambda \circ g^q(\mathcal{X}, x)$.

Then we can formulate the following result.

Theorem 9. Let the hypothesis of Theorem 6 hold, let Assumption 5 hold and consider the nonlinear MPC problem (5) at time $k \in \mathbb{N}$ with prediction horizon $N(k) \geq 2$ and the additional constraint $x(N(k)|k) \in \mathbb{X}_T$, where $\mathbb{X}_T \subseteq \mathbb{X}$ is a terminal set satisfying Assumption 8 for a state–feedback control law $\kappa(x)$. Then the following properties hold:

- If the nonlinear MPC problem (5) with the additional terminal constraint is feasible at time k, then the same problem is feasible at time $k+1$ for any $N(k+1) \geq N(k)$;
- Assume there exists $\sigma \in \mathcal{K}$ such that
$$l(x, \kappa(x)) \leq \sigma \circ g^q(\mathcal{X}, x), \quad \forall x \in \mathbb{X}_T. \tag{12}$$
Then there exists a $N(k+1) \geq N(k)$ such that the negative supply condition (10) holds at time $k+1$.

Proof. Since $x^*(N(k)|k) \in \mathbb{X}_T$, we can construct a feasible shifted sequence

$$\begin{aligned} &\{u^*(1|k), \ldots, u^*(N(k)-1|k), \\ &\quad \kappa(x^*(N(k)|k)), \ldots, \kappa(x(N(k)+N_T|k))\} \end{aligned} \tag{13}$$

for any $N_T \geq 1$, where

$$\begin{aligned} &x(N(k)+N_T|k) = \\ &f(x(N(k)+N_T-1|k), \kappa(x(N(k)+N_T-1|k))) \in \mathbb{X}_T \end{aligned}$$

for all $N_T \geq 1$. Hence, the MPC optimization problem with the terminal constraint added is feasible at time $k+1$ for any $N(k+1) \geq N(k)$.

Moreover, $x^*(N(k)|k) \in \mathbb{X}_T$ implies $g^q(\mathcal{X}, x^*(N(k)|k)) \leq c$. Hence, from inequality (12) it follows that there exists a $N_T \geq 1$ such that

$$\begin{aligned} &l(x(N(k)+N_T|k), \kappa(x(N(k)+N_T|k))) \\ &\leq \sigma \circ g^q(\mathcal{X}, x(N(k)+N_T|k)) \\ &\leq \sigma \circ \lambda^{N_T}(c) \\ &\leq (\mathrm{id} - \rho) \circ l(x(k), u^*(0|k)), \end{aligned}$$

for any $x(k) \neq 0$, which implies condition (10) holds for $N \geq N(k) + N_T$. Above we have used that $\lambda < \mathrm{id}$ and $\sigma \in \mathcal{K}$. Hence, by selecting $N(k+1) = N(k) + N_T - 1$ and time $k+1$ we have that the dissipation inequality (7) holds with $\tilde{V}(x(k))$ instead of $V(x(k))$ corresponding to the extended sequence (13) and with the supply $s(x(k), u(k))$ satisfying (10). □

The above result provides a mechanism for adapting the prediction horizon online such that recursive feasibility

and convergence of the optimal nonlinear MPC cost function are guaranteed (assuming $N(k)$ converges to a constant value in finite time), for an arbitrary positive definite cost function. An alternative way of guaranteeing convergence with a time–varying $N(k)$ is to employ a contractive terminal constraint, as proposed in (Limon et al., 2006), or to resort to time–varying Lyapunov functions. The details of these approaches will be presented in an extended paper.

Observe that the condition (12) is easily satisfied if the terminal set is such that its Minkowski functional is a norm, e.g., an ellipsoidal set. Indeed, since typically $l(x, \kappa(x))$ is upper bounded by a $\mathcal{K}$–function of x in the terminal set, condition (12) holds due to equivalence of norms. For systematic computation of ellipsoidal contractive sets for discrete–time nonlinear systems we refer to the sNMPC toolbox (Eyüboğlu and Lazar, 2022).

5. PRACTICAL STABILIZING NMPC DESIGN

As a conclusion to this extended abstract we propose two practical methods for designing stabilizing NMPC algorithms with arbitrary positive definite stage cost functions.

Time–varying horizon NMPC

- At time $k \in \mathbb{N}$, given $x(k)$, $N(k)$, $l(\cdot,\cdot)$, $\mathbb{X}_T$ and $\kappa(\cdot)$, $\mathbb{X}$ and $\mathbb{U}$ solve the optimization problem (assuming it is feasible):

$$\min_{\mathbf{u}(k)} \sum_{i=0}^{N(k)-1} l(x(i|k), u(i|k))$$

subject to constraints:

$$x(i+1|k) = f(x(i|k), u(i|k)), \quad \forall i \in \mathbb{N}_{[0,N(k)-1]},$$
$$(x(i+1|k), u(i|k)) \in \mathbb{X} \times \mathbb{U}, \quad \forall i \in \mathbb{N}_{[0,N(k)-1]},$$
$$x(N(k)|k) \in \mathbb{X}_T.$$

- Compute $l(x^*(N(k)|k), \kappa(x^*(N(k)|k)))$, where $\kappa(\cdot)$ is the state–feedback control law that renders $\mathbb{X}_T$ contractive and check if

$$l(x^*(N(k)|k), \kappa(x^*(N(k)|k))) < l(x(k), u^*(0|k)).$$

  If the above inequality holds, apply $u^*(0|k)$, set $N(k+1) = N(k)$ and repeat for $k+1$.
- Else, compute $l(x(N(k)+N_T|k), \kappa(x(N(k)+N_T|k)))$ for $N_T = 1, 2, \ldots$ using the dynamics $f(x, \kappa(x))$ until

$$l(x(N(k)+N_T|k), \kappa(x(N(k)+N_T|k))) < l(x(k), u^*(0|k)).$$

  Then apply $u^*(0|k)$, set $N(k+1) = N(k) + N_T$ and repeat for $k+1$.

If the terminal set and local controller $\kappa(\cdot)$ are not available/used, the above algorithm can still be implemented by computing the additional control inputs online, solving additional optimization problems, as suggested in (Lazar, 2021).

Alternatively, the following NMPC implementation can be used, which explicitly enforces the stage cost constraint and utilizes a slack variable s, to minimize the supply function induced by the stage cost. Simulation examples will be provided during the presentation and in a corresponding article.

NMPC with flexible supply constraint

- At time $k \in \mathbb{N}$, given $x(k)$, N, $l(\cdot,\cdot)$, $\mathbb{X}_T$ and $\kappa(\cdot)$, $\mathbb{X}$, $\mathbb{U}$, $\lambda > 0$ and $\varepsilon \in \mathbb{N}_{(0,1)}$ solve the optimization problem (assuming it is feasible):

$$\min_{\mathbf{u}(k), s(k)} \sum_{i=0}^{N-1} l(x(i|k), u(i|k)) + \lambda s(k)$$

subject to constraints:

$$l(x(N|k), \kappa(x(N|k))) - \varepsilon l(x(k), u(0|k)) \leq s(k)$$
$$x(i+1|k) = f(x(i|k), u(i|k)), \quad \forall i \in \mathbb{N}_{[0,N-1]},$$
$$(x(i+1|k), u(i|k)) \in \mathbb{X} \times \mathbb{U}, \quad \forall i \in \mathbb{N}_{[0,N-1]},$$
$$x(N|k) \in \mathbb{X}_T.$$

REFERENCES

Bemporad, A., Chisci, L., and Mosca, E. (1994). On the Stabilizing Property of SIORHC. *Automatica*, 30(12), 2013 – 2015.

Blanchini, F., Miani, S., Blanchini, F., and Miani, S. (2015). *Set-Theoretic Methods in Control*. Springer International Publishing.

Boccia, A., Grüne, L., and Worthmann, K. (2014). Stability and feasibility of state constrained MPC without stabilizing terminal constraints. *Systems & Control Letters*, 72, 14 – 21.

Eyüboğlu, M. and Lazar, M. (2022). sNMPC: A Matlab Toolbox for Computing Stabilizing Terminal Costs and Sets. In *25th International Symposium on Mathematical Theory of Networks and Systems (MTNS)*. Bayreuth, Germany. URL `https://github.com/mlazar04/sNMPC`.

Grüne, L. (2012). NMPC without Terminal Constraints. In *Proceedings of 4th IFAC Conference on Nonlinear Model Predictive Control*, 1–13. Noordwijkerhout, Netherlands.

Lazar, M. (2006). *Model Predictive Control of Hybrid Systems*. Eindhoven University of Technology, PhD Thesis.

Lazar, M. (2021). A dissipativity-based framework for analyzing stability of predictive controllers. In *7th IFAC Conference on Nonlinear Model Predictive Control*. Bratislava, Slovakia.

Limon, D., Alamo, T., and Camacho, E. (2005). Enlarging the domain of attraction of MPC controllers. *Automatica*, 41(4), 629–635.

Limon, D., Alamo, T., Salas, F., and Camacho, E.F. (2006). On the Stability of Constrained MPC Without Terminal Constraint. *IEEE Transactions on Automatic Control*, 51(5), 832 – 836.

Mayne, D.Q. (2001). Control of Constrained Dynamic Systems. *European Journal of Control*, 7, 87 – 99.

Mayne, D.Q. (2013). An apologia for stabilising terminal conditions in model predictive control. *International Journal of Control*, 86(11), 2090 – 2095.

Mayne, D.Q. (2014). Model predictive control: Recent developments and future promise. *Automatica*, 50, 2967 – 2986.

Rawlings, J.B., Mayne, D.Q., and Diehl, M.M. (2017). *Model Predictive Control: Theory, Computation, and Design, 2nd Edition*. Nob Hill Publishing.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Convex Approach to Data-Driven Stochastic Optimal Control using Linear Operator Theory ⋆

Umesh Vaidya

* *Department of Mechanical Engineering, Clemson University, Clemson SC 29631 (email: uvaidya@clemson.edu)*

EXTENDED ABSTRACT

The stochastic optimal control problem (SOCP) is a cornerstone of systems and control theory. This problem has received renewed attention with the growing interest in data-driven analytics and control with applications ranging from vehicle autonomy, robotics, transportation networks, power grid, security, and advanced manufacturing. The SOCP is also at the heart of Reinforcement learning (RL), and various algorithms are developed for the data-driven approximation of its solution. For a system in continuous time with continuous state space and control, the solution to SOCP essentially boils down to solving a Hamilton Jacobi Bellman (HJB) equation, which is a nonlinear partial differential equation. In discrete-time, SOCP involves solving the Bellman equation using the principle of dynamic programming. Thus, the Bellman equation can be viewed as the discrete-time counterpart of the continuous-time HJB equation. Given the nonlinear nature of the HJB equation, one of the popular approaches to solving the HJB equation is the iterative approach. In this work, we propose an alternate method for solving SOCP based on the convex formulation of the problem in the dual space of densities. We provide a data-driven solution to the SOCP over an infinite time horizon with continuous-time system dynamics. The dual approach leads to a convex infinite-dimensional optimization problem to be solved for the SOCP. Unlike iterative algorithms for solving HJB equation in the primal domain, the convex formulation in the dual space lends itself to a single-shot approach for solving SOCP. We use a linear operator theoretic framework involving P-F and Koopman operators to provide a novel perspective on the SOCP problem and the computation of its solution using data. We show that the traditional primal formulation of SOCP involving the HJB equation is closely tied to the Koopman operator. Furthermore, the dual convex formulation of the SOCP can be understood naturally through the lenses of duality between the Koopman and P-F operator.

We provide a convex formulation to the SOCP using the linear operator theoretic framework involving P-F and Koopman operators. The linear P-F and Koopman operators are dual and provide a linear lifting of nonlinear system dynamics in the space of density and function (observables), respectively. The dynamical system theory inspires the results, as the duality in SOCP is discovered through duality between the P-F and Koopman operators. The SOCP problem is formulated using P-F operator-based lifting of control system dynamics in the dual density space. This dual approach leads to the infinite-dimensional convex optimization-based formulation of the SOCP. We provide a computational framework based on the data-driven approximation of the P-F operator for the data-driven stochastic optimal control design. The convex formulation of SOCP is made possible by exploiting the P-F operator's linearity, positivity, and Markov properties. Furthermore, we show that the hard constraints on the control input and the state can also be written convexly in the dual formulation. The state constraints will include safety or obstacle avoidance constraints. On the other hand, we establish a connection between the Koopman operator and the HJB equation. This connection allows us to develop a numerical algorithm for the data-driven solution of the HJB equation based on Koopman theory. In particular, we provide an iterative algorithm based on a data-driven approximation of the Koopman operator for solving the SOCP problem in the primal domain. This new algorithm is reminiscent of the generalized policy iteration (GPI) algorithm in RL, and we call it Koopman policy iteration (KPI). Moreover, the interpretation of GPI using the Koopman theory opens up the possibility of exploiting the rich spectral theory of the Koopman operator for data-driven control. It is important to emphasize that the existing iterative algorithm for solving the HJB equation, including our proposed Koopman-based approach, requires an initial control policy to be stabilizing. However, designing stabilizing controller for a stochastic nonlinear system is far from a trivial problem. Our proposed dual approach to SOCP does not suffer from this drawback. The convex optimization problem in our dual framework can be solved as a single shot problem, where almost everywhere, stochastic stabilizability arises as a constraint of this optimization problem. So the data-driven stochastic stabilization will emerge as the particular case of the main result on SOCP.

⋆ Financial support from of NSF CPS award 1932458 and NSF 2031573 is greatly acknowledged.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Algebraic and Path-Based Conditions for Local Network Identifiability

A. Legat, J. M. Hendrickx *

* *ICTEAM Institute, UCLouvain, B-1348 Louvain-la-Neuve, Belgium.*
(e-mail: { antoine.legat, julien.hendrickx } @uclouvain.be).

Abstract: This work focuses on the generic identifiability of dynamical networks with partial excitation and measurement: a set of nodes are interconnected by transfer functions according to a known topology, some nodes are excited, some are measured, and only a part of the transfer functions are known. Our goal is to determine whether the unknown transfer functions can be generically recovered based on the input-output data collected from the excited and measured nodes. Introducing the notion of generic local identifiability, we derive a necessary and sufficient algebraic condition, which can be checked efficiently by rank computation. Another notion, generic decoupled identifiability, allows to reflect on a larger network which decouples excitations and measurements. This yields a necessary path-based condition, and a sufficient one.

Keywords: System Identification, Networked Control Systems, Linear Systems.

1. INTRODUCTION

The goal of this work is to recover the local dynamics from the global input-output behavior of a networked system and the knowledge of the network topology.

We consider the identifiability of a network matrix $G(q)$, where the network is made up of n node signals $w(t)$, external excitation signals $r(t)$, measured nodes $y(t)$ and noise $v_1(t), v_2(t)$ related to each other by:

$$\begin{aligned} w(t) &= G(q)w(t) + Br(t) + v_1(t) \\ y(t) &= Cw(t) + v_2(t), \end{aligned} \tag{1}$$

where matrices B and C are binary selections defining respectively the n_B excited nodes and n_C measured nodes, forming sets $\mathcal{B}$ and $\mathcal{C}$ respectively. The nonzero entries of the network matrix $G(q)$ define the network topology: some of them are known and collected in $G^0(q)$, and the others are the unknowns to identify, collected in $G^\Delta(q)$, such that $G(q) = G^0(q) + G^\Delta(q)$.

We assume that the global relation between the excitations r and measurements y has been identified, and that the structure of $G(q)$ is known. From this knowledge, we aim at recovering the unknown entries of $G(q)$.

Networked systems have recently been the object of a significant research effort. The independent identification of all local dynamics in a networked system would require exciting and measuring every single node of the network, which is costly and often impractical. We therefore assume here that we excite and measure different subsets of nodes, and are able to identify the global input-output dynamics going from the excited nodes to the measured nodes.

A recent approach employs graph-theoretical tools to derive identifiability conditions on the graph of the network. Using this approach, Hendrickx et al. (2018); Cheng et al. (2019) address the particular case where all nodes are excited. In the general case of partial measurement *and* excitation, Shi et al. (2021) derive identifiability conditions while exploiting unmeasured noise. In this work, we consider partial measurement and excitation and we introduce the notions of generic local and decoupled identifiability, for which we derive algebraic and path-based conditions.

Assumptions: Consistently with previous works, we assume that $(I - G(q))^{-1}$ is proper and stable, and we consider a single frequency z, so that all transfer functions are modeled simply by a complex value, and the matrices G and $T(G) = (I-G)^{-1}$ are complex matrices rather than matrices of transfer functions.

Genericity: Besides, we say that a network is *generically* identifiable if it is identifiable at all G with same known and zero entries, except possibly those lying on a set of dimension lower than the number of unknown edges. In the remainder of this paper, we say that a property is *generic* if it either holds (i) for *almost all* variables, i.e for all variables except possibly those lying on a lower-dimensional set, or (ii) for no variable.

2. LOCAL IDENTIFIABILITY

First, we remind a notion of identifiability amenable to linear analysis: local identifiability, which corresponds to identifiability provided that $\tilde{G}$ is sufficiently close to G.

Definition 1. The network is *locally identifiable* at G from excitations $\mathcal{B}$ and measurements $\mathcal{C}$ if there exists $\epsilon > 0$ such that for any $\tilde{G}$ with same zero and known entries as G satisfying $||\tilde{G} - G|| < \epsilon$, there holds

$$C\,T(\tilde{G})\,B = C\,T(G)\,B \Rightarrow \tilde{G}^\Delta = G^\Delta, \tag{2}$$

where $\tilde{G}^\Delta$ collects only the entries of $\tilde{G}$ corresponding to unknown edges, just as G^Δ does for G.

Local identifiability is a generic property of the network, and is a necessary condition for identifiability. It is *a priori* a weaker notion, yet no example locally identifiable but not globally identifiable is known to the authors.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2.1 Necessary and sufficient algebraic condition

Linearizing (2) yields a necessary and sufficient condition.

Proposition 2. Exactly one of the two following holds:

(i) the network is generically locally identifiable and for almost all G, there holds
$$C\,T(G)\,\Delta\,T(G)\,B = 0 \Rightarrow \Delta = 0, \tag{3}$$
for all Δ of same dimensions and zero entries as G^{Δ}.
(ii) the network is never locally identifiable and there is no G for which (3) holds $\forall\ \Delta$ with same 0s as G^{Δ}.

Moreover, (3) holds if and only if matrix $K(G)$ is full rank. Expression of $K(G)$ can be found in Legat and Hendrickx (2020), along with an algorithm that efficiently determines generic local identifiability by rank computation of $K(G)$.

The algebraic condition of Proposition 2 allows rapidly testing local identifiability for any given network, but we aim at finding a combinatorial characterization for generic identifiability, that is expressed purely in terms of path-based properties, akin to what was done in the full excitation case e.g. in Hendrickx et al. (2018). This spurs the need for a new notion of identifiability.

3. DECOUPLED IDENTIFIABILITY

We consider a more general notion than local identifiability. It is essentially inspired from (3), where the left and right $T(G)$ are no more equal, hence the name *decoupled*.

Definition 3. A network is *decoupled-identifiable* at (G, G'), with G and G' sharing the same zero and known entries, if for all Δ with same zeros as G^{Δ}, there holds:
$$C\,T(G)\,\Delta\,T(G')\,B = 0 \Rightarrow \Delta = 0. \tag{4}$$

Similarly to local identifiability, decoupled identifiability is a generic property: either it holds for almost all (G, G'), or for no (G, G'). Besides, generic decoupled identifiability is necessary for generic local identifiability, which is itself necessary for generic identifiability. Hence, necessary conditions obtained for generic decoupled identifiability apply to (generic) (local) identifiability as well.

In addition, generic decoupled identifiability of a network is equivalent to generic identifiability of a larger network, constructed by duplicating the initial network, exciting one copy, measuring the other one and adding the unknown edges in the middle. This larger network is called *decoupled network* and is defined in Legat and Hendrickx (2021).

3.1 Necessary path-based condition, and sufficient

Decoupled identifiability allows to reflect on the decoupled network, on which for each unknown edge, one can route a path that starts at an excitation and ends at a measurement. Building on Proposition 2, this approach leads to Theorem 4, which gives a necessary path-condition and a sufficent one, for generic decoupled identifiability.

For ease of presentation, we consider the case where there are exactly $n_B n_C$ unknown edges, i.e. as many as the number of (excitation, measurement) pairs. Transfer functions are referred to as edges, and an *assignation* σ is a function that assigns to each unknown edge a pair (excitation, measure). We say that σ is *connected* if for each unknown edge there is a path from its assigned excitation to its assigned measurement, in which the unknown edge is included.

Theorem 4. If a network is generically decoupled-identifiable, then there is at least one assignation σ such that:

(a) n_C unknown edges are assigned to each excitation
(b) n_B unknown edges are assigned to each measurement
(c) σ is connected
(d) for each excitation b, there are n_C vertex-disjoint paths between the edges assigned to b and measurements $\mathcal{C}$.
(e) for each measurement c, there are n_B vertex-disjoint paths between edges assigned to c and excitations $\mathcal{B}$.

If there is only one such assignation, then this condition is also sufficient for generic decoupled identifiability.

4. DISCUSSION

In Legat and Hendrickx (2021), the intuition is given that a stronger version of Theorem 4 could exist, which would extend path-based conditions of Hendrickx et al. (2018). Specific counter-examples allow to narrow the search for such stronger conditions and a low-level approach gives a more clear understanding of the problem.

Besides, we remind that the necessary condition of Theorem 4 extends to (generic) (local) identifiability, and that no counter-example to sufficiency of generic decoupled and local identifiability is known to the authors.

5. CONCLUSION

We introduced the notion of local generic identifiability, which excludes situations where the set of solutions is discrete. Linearizing this notion gave a necessary and sufficient algebraic condition, which can be checked efficiently by rank computation.

Another new notion, decoupled identifiability, allowed us to reflect on a larger graph that highlights paths between excitations and measurements. This approach led to a necessary path-based condition, and a sufficient one.

Eventually, potential stronger path-based conditions are discussed, supported by specific counter-examples and a low-level approach.

REFERENCES

Cheng, X., Shi, S., and Van den Hof, P.M. (2019). Allocation of excitation signals for generic identifiability of dynamic networks. In *Proceedings of the IEEE CDC*.

Hendrickx, J.M., Gevers, M., and Bazanella, A.S. (2018). Identifiability of dynamical nets with partial node measurements. *IEEE Transactions on Automatic Control*.

Legat, A. and Hendrickx, J.M. (2020). Local network identifiability with partial excitation and measurement. In *2020 59th IEEE CDC*.

Legat, A. and Hendrickx, J.M. (2021). Path-based conditions for local network identifiability–full version. In *2021 60th IEEE CDC*.

Shi, S., Cheng, X., and Van den Hof, P.M. (2021). Single module identifiability in linear dynamic networks with partial excitation and measurement. *IEEE TAC*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Nash equilibria of the pay-as-bid auction with K-Lipschitz supply functions

EXTENDED ABSTRACT

M. Vanelli * **G. Como** * **F. Fagnani** *

* *Department of Mathematical Sciences, Politecnico di Torino*
(e-mail: {martina.vanelli,giacomo.como, fabio.fagnani}@polito.it).

Abstract: We model a system made of n asymmetric firms participating in a market in which each firm chooses as its strategy a supply function relating its quantity to its price. Such strategy (Supply function equilibrium) is a generalization of models where firms can either set a fixed quantity (Cournot model) or set a fixed price (Bertrand model). Our goal is to study the pay-as-bid auction in this setting. Under the assumption of K-Lipschitz supply functions, we were capable of determining existence and characterization of Nash equilibria of the game.

Keywords: Supply function equilibrium, Pay-as-bid auction, Discriminatory price auction, Oligopoly, Electricity market

1. INTRODUCTION

The progressive liberalization of electricity markets motivates the need to develop realistic and robust models for the analysis of the strategic bidding problem (Ventosa et al. (2005)). Pricing rules in oligopolistic wholesale electricity auctions are mainly two: the uniform price rule and the pay-as-bid rule (Rassenti et al. (2003), Fabra et al. (2006)). In a uniform price auction, electricity is paid/sold at the market-clearing price, regardless of the offers that bidders actually made. On the other hand, in the pay-as-bid auction (also called discriminatory price auction), the remuneration is the bid price. Uniform price auctions are usually employed in day-ahead markets, while ancillary services markets sometimes adopt pay-as-bid remunerations (e.g, see Müsgens et al. (2014) and Gestore dei Mercati Energetici (GME)).

From a game-theoretic point of view, appropriate models for studying wholesale markets for electricity are Supply Function Equilibrium (SFE) models. With this approach, instead of setting their price bids (Bertrand) or quantities (Cournot), see Mas-Colell et al. (1995), firms bid their choices of supply functions and the predicted outcome is a Nash equilibrium of the game. SFE models were first introduced by Klemperer and Meyer (1989), and then applied to electricity markets by Green and Newbery (1992)). Empirical studies of strategic bidding suggest that the SFE model provides a good approximation of the behavior of large producers (Hortaçsu and Puller (2008); Sioshansi and Oren (2007)).

There is a vast amount of literature directed to the study of SFE outcomes in uniform-price auctions (e.g., David (1993), Baldick and Hogan (2001), Anderson and Philpott (2002), Baldick et al. (2004), Holmberg and Newbery (2010)), but less clear is the behavior of SFE models when discriminatory prices are considered. Our focus is on existence and characterization of Nash equilibria in supply functions with the pay-as-bid remuneration and asymmetric firms. We determine conditions on the strategy space under which existence is guaranteed and best responses can be characterized. Our work is related to Holmberg (2009), where uncertainty is considered. The authors determine conditions on the hazard rate of the demand distribution to ensure existence of Nash equilibria which is in general not guaranteed. We instead study the problem from a deterministic perspective with the goal of determining a tractable model. Although different in the purpose, it is relevant to mention Genc (2009), where supply function equilibria game models are compared for uniform-price and pay-as-bid auctions.

2. MODEL

2.1 Definition of the game

In the general setting, we shall consider:

- a player set $\mathcal{N} = \{1, \dots, n\}$ made of n firms equipped with non-decreasing *cost functions* $C_i(q)$, for $i \in \mathcal{N}$, where q denotes the sold quantity. We assume that C_i is twice continuously differentiable and such that $C_i' \geq 0$ and $C_i'' \geq 0$ for every $i \in \mathcal{N}$. The assumption of convex costs is standard in literature (see Klemperer and Meyer (1989), Green and Newbery (1992));
- a strictly decreasing industry *demand function* $D(p)$, which retuns the aggregate quantity that consumers are willing to buy at a (maximum) unit price p. We define $\hat{p}$ as the price such that $D(\hat{p}) = 0$. We assume that $D \in C^2$ and such that $D' < 0$ and $D'' \leq 0$.

The *strategy* of firm $i \in \mathcal{N}$ is a supply function belonging to an arbitrary subset of the set of non-decreasing continuous functions passing through the origin, i.e.,

$$\mathcal{F} = \{S \in C^0([0, \hat{p}])\,,\ S_i(0) = 0\,,\ S \text{ non-decreasing}\}\,. \quad (1)$$

The supply function S_i returns the quantity $q = S_i(p)$ that the firm is willing to produce at (minimum) unit price p. The *strategy configuration* of the auction combines all

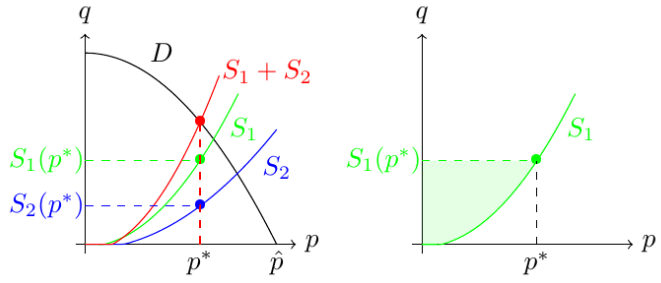


Fig. 1. The equilibrium marginal price (on the left) and the pay-as-bid remuneration (on the right).

strategies, that is, $\mathbf{S} = (S_1, \dots, S_n)$. For a firm $i \in \mathcal{N}$ and a strategy configuration $\mathbf{S}$, we shall refer to the other firms' strategies with $S_{-i} = \{S_j\}_{j \neq i}$.

Given a demand D and a strategy configuration $\mathbf{S}$, the *equilibrium marginal price* is determined as the price that matches total demand and total supply, that is, $p^* \in [0, \hat{p}]$ satisfying

$$D(p^*) = \sum_{i=1}^{n} S_i(p^*)\,. \tag{2}$$

We remark that p^* exists unique in $[0, \hat{p}]$ under the assumptions of a strictly decreasing continuous demand function and increasing continuous supply functions satisfying $S_i(0) = 0$ for all i. The equilibrium marginal price determines the total quantity that will be sold by each firm in the auction, that is, $q_i^* = S_i(p^*)$ for every $i \in \mathcal{N}$. An example of equilibrium marginal price is depicted on the left of Fig.1.

We define the following class of games based on the pay-as-bid remuneration. For a given $\mathcal{A} \subseteq \mathcal{F}$, the *pay-as-bid* (PAB) *auction* is a game with player set $\mathcal{N}$, strategy space $\mathcal{A}$ and utilities, for every $i \in \mathcal{N}$,

$$u_i(S_i, S_{-i}) := p^* S_i(p^*) - \int_0^{p^*} S_i(p)\,\mathrm{d}p - C_i(S_i(p^*))\,, \tag{3}$$

where $p^* := p^*(S_i, S_{-i})$ is the equilibrium marginal price satisfying (2). We shall denote the PAB auction with $\mathcal{U} = (\mathcal{N}, \mathcal{A}, \{u_i\}_{i \in \mathcal{N}})$.

In words, firm i sells $S_i(p^*)$ at the bid price and the final utility is given by the total revenue minus the production cost. We remark that the integral term is the one that makes the difference between the uniform-price auction and the pay-as-bid one. Indeed, notice that, when S_i is differentiable,

$$\int_0^{p^*} pS_i'(p)\,\mathrm{d}p = p^* S_i(p^*) - \int_0^{p^*} S_i(p)\,\mathrm{d}p\,.$$

If the reward were only $p^* S_i(p^*)$, we would be dealing with a uniform price auction, thus obtaining the model in Klemperer and Meyer (1989). Observe that, if S_i is invertible, the total revenue in the pay-as-bid auction equals the integral from 0 to $S_i(p^*)$ of the inverse of S_i, that is, the price function $P_i(q) := S_i^{-1}(q)$ of producer i. The price function assigns to each quantity the marginal price at which firms are willing to sell such quantity for. Therefore, its integral from 0 to q_i^* determines the total pay-as-bid remuneration for firm i for a quantity q_i^*. By considering the formula in (3), we do not need to make any assumption on S_i.

An example of remuneration of the pay-as-bid auction is depicted on the right of Figure 1. When the supply function is S_1 and the equilibrium marginal price is p^*, the total revenue for firm 1 coincides with the green area (the utility is then given by revenue minus costs).

Throughout the analysis, we shall focus on existence and characterization of Nash equilibria of the pay-as-bid auction: an action configuration $\mathbf{S}$ is a (pure strategy) *Nash equilibrium* if, for every $i \in \mathcal{N}$, S_i maximizes the utility given the other firms' strategies. Let $S_{-i} \in \mathcal{A}^{\mathcal{N}\setminus\{i\}}$. We shall refer to the set

$$\mathcal{B}_i(S_{-i}) = \operatorname*{argmax}_{S_i \in \mathcal{A}} u_i(S_i, S_{-i})$$

as the *best response* of firm i to S_{-i}. Then, $\mathbf{S}$ is a Nash equilibrium if and only if $S_i \in \mathcal{B}_i(S_{-i})$ for every $i \in \mathcal{N}$.

2.2 Remarks on the strategy space

In this section, in order to provide the motivation for our work, we shall discuss existence of Nash equilibria in the pay-as-bid auction depending on the choice of the strategy space $\mathcal{A}$.

Let us first observe that, when the supply functions can be generic non-increasing continuous functions, that is, when $\mathcal{A} = \mathcal{F}$ as in (1), the PAB auction admits no Nash equilibria. More precisely, we shall prove that the best-response does not exist.

Proposition 1. Consider the PAB auction with strategy space $\mathcal{A} = \mathcal{F}$ as in (1). Then, for every $i \in \mathcal{N}$ and $S_{-i} \in \mathcal{A}^{\mathcal{N}\setminus\{i\}}$, $\mathcal{B}_i(S_{-i}) = \emptyset$.

Proof. We shall prove that for every feasible supply function $S_i \in \mathcal{A}$, there exists another feasible supply function $\tilde{S}_i \in \mathcal{A}$ yielding to the same equilibrium price and a higher utility. Formally, let S_i be any non-decreasing continuous function yielding to an equilibrium price p^*. We shall then define

$$\tilde{S}_i(p) := S_i\left(\frac{p^2}{p^*}\right)\,.$$

Observe that $\tilde{S}_i(0) = S_i(0)$ and $\tilde{S}_i(p^*) = S_i(p^*)$. Also $S_i(p) \geq \tilde{S}_i(p)$ for all $p \in [0, p^*]$ and, more precisely, $S_i(p) > \tilde{S}_i(p)$ for every $p \in (0, p^*)$ such that $S_i(p) > 0$. Then,

$$u_i(\tilde{S}_i, S_j) > u_i(S_i, S_j)\,.$$

This concludes the proof.

Remark 1. The proof of Proposition 1 suggests that best responses would exist if one could use step functions. However, enlarging the strategy space to discontinuous functions would lead to a number of different technical difficulties. For instance, one has to solve some technical problems in the definition of the game. Indeed, the existence of a unique marginal equilibrium price p^* is not anymore guaranteed. Anyway, even when we technically solve such problem, Nash equilibria might fail to exist.

Thus, we observed that, in the general settings, Nash equilibria might fail to exist. This gives the motivation for the K-Lipschitz assumption, which guarantees existence and characterization of Nash equilibria.

3. RESULTS

3.1 The Lipschitzianity assumption

As previously observed, one of the main issues is that, without any particular restriction on the strategy space, the best response is a step function and existence of Nash equilibria is not guaranteed. Accordingly, the best response does not exist when considering generic continuous supply functions. The problem is solved when considering K-Lipschitz supply functions, for a fixed $K > 0$. Under this assumption, not only best responses do exist, but it is rather simple to determine their structure. As we shall observe, this assumption drastically reduces the dimention of the strategy space.

Let $K > 0$. We recall that a function $S : [0, \hat{p}] \to [0, \infty)$ is *K-Lipschitz* if

$$|S(x) - S(y)| \leq K|x - y|\,, \quad \forall x, y \in [0, \hat{p}]\,,\, x \neq y\,.$$

Let us then define

$$\mathcal{A}_K := \{S \in \mathcal{F}\,, S \text{ is } K\text{-Lipschitz}\}\,. \tag{4}$$

Under the assumption of K-Lipschitz supply functions, that is, $\mathcal{A} = \mathcal{A}_K$, we can characterize best response functions, as showed in the following lemma.

Lemma 1. (Affine best-response). Consider the PAB auction $\mathcal{U}$ with strategy space $\mathcal{A} = \mathcal{A}_K$. Then, for every $i \in \mathcal{N}$ and $S_{-i} \in \mathcal{A}^{\mathcal{N}\setminus\{i\}}$,

$$S_i \in \mathcal{B}_i(S_{-i}) \quad \Rightarrow \quad S_i(p) = K[p - p_i]_+ \tag{5}$$

for some $p_i \in [0, \hat{p}]$.

Proof. Consider a generic function $S_i \in \mathcal{A}_K$ and let p^* denote the equilibrium price such that $D(p^*) = \sum_{j=1}^n S_j(p^*)$. We shall prove the statement by construction, that is, we provide a strategy $\tilde{S}_i \in \mathcal{A}_K$ of the form in (5) that gives a greater or equal utility than S_i. More precisely, we construct $\tilde{S}_i$ of the form in (5) that passes through the same p^* as S_i, that is, $\tilde{S}_i(p) = K[p - p_i]_+$ with

$$p_i := p^* - \frac{S_i(p^*)}{K}\,.$$

The K-Lipschitzianity of S_i combined with the assumption that $S_i(0) = 0$ guarantees that $p^* - \frac{S_i(p^*)}{K} \geq 0$ and thus ensures the existence of $p_i \geq 0$. An example is shown in Fig. 2 for S_1 and S_2 as in Fig. 1 and $K = 3$. We remark that $\tilde{S}_i \in \mathcal{A}_K$.

We shall now prove that $S_i(p) \geq \tilde{S}_i(p)$ for all $p \in [0, p^*]$. First, notice that, for $p \in [0, p_i]$, the inequality is trivial since $S_i(p) \geq \tilde{S}_i(p) = 0$. Let $p \in (p_i, p^*]$. Then, the inequality is satisfied as

$$\begin{aligned} S_i(p^*) - S_i(p) = |S_i(p^*) - S_i(p)| &\overset{(1)}{\leq} K|p^* - p| = K(p^* - p) \\ &= K(p^* - p_i) - K(p - p_i) \\ &= \tilde{S}_i(p^*) - \tilde{S}_i(p) \\ &= S_i(p^*) - \tilde{S}_i(p) \end{aligned}$$

where (1) is guaranteed by the K-Lipschitz property of S_i.

To sum up, we observed that $S_i(p) \geq \tilde{S}_i(p)$ for all $p \in [0, p^*]$ and $S_i(p^*) = \tilde{S}_i(p^*)$. Therefore,

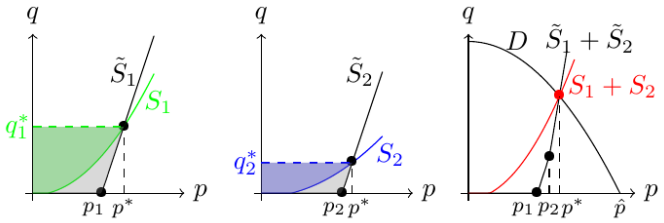


Fig. 2. Explanation of Lemma 1 (see Remark 2).

$$\begin{aligned} u_i(S_i, S_{-i}) &= p^* S_i(p^*) - \int_0^{p^*} S_i(p)dp - C_i(S_i(p^*)) \\ &\leq p^* \tilde{S}_i(p^*) - \int_0^{p^*} \tilde{S}_i(p)dp - C_i(\tilde{S}_i(p^*)) \\ &= u_i(\tilde{S}_i, S_{-i})\,. \end{aligned}$$

More precisely, it holds that, $\forall\, p \in [0, p^*]$,

$$\int_0^{p^*} S_i(p)\,\mathrm{d}p = \int_0^{p^*} \tilde{S}_i(p)\,\mathrm{d}p \quad \Leftrightarrow \quad S_i(p) = \tilde{S}_i(p)\,,$$

which implies that $u_i(S_j, S_i) = u_i(S_j, \tilde{S}_i)$ if and only if $S_i \equiv \tilde{S}_i$. This concludes the proof.

Remark 2. Lemma 1 is illustrated in Fig. 2. Consider two generic supply functions S_1 and S_2 as in Fig.1. Notice that when playing $\tilde{S}_1(p) = K[p - p_1]_+$ for p_1 as in figure, firm 1 receives a higher utility than the one obtained by playing S_1. Indeed, the remuneration increases (colored areas), while the equilibrium price does not change, thus yielding to the same sold quantity. The same happens for firm 2 when playing $\tilde{S}_2(p) = K[p - p_2]_+$ instead of S_2. Then, for any supply S_i, it is possible to construct another supply $\tilde{S}_i$ yielding to a higher utility. Thus, best responses must have such form.

Lemma 1 yields a fundamental simplification in our problem. Indeed, according to (5), every Nash equilibrium $\mathbf{S}^*$ necessarily exhibits the form

$$S_i^*(p) = K[p_i - p^*]_+$$

for suitable values $p_i \in [0, \hat{p}]$. In particular, this yields a complexity reduction from an infinite-dimensional strategy space to a finite-dimensional one. In order to find Nash equilibria, we can indeed further restrict the strategy space to considering just functions as in (5), which are parametrized by just one parameter, that is, $p_i \in [0, \hat{p}]$, for $i \in \mathcal{N}$. We can then define a restricted game $\mathcal{U}^r = (\mathcal{N}, \mathcal{A}^r, \{u_i^r\}_{i\in\mathcal{N}})$, where $\mathcal{A}^r = [0, \hat{p}]$ and the utilities are functions of p_i and $p_{-i} = \{p_j\}_{j\neq i}$, that is, for $i \in \mathcal{N}$,

$$\begin{aligned} u_i^r(p_i, p_{-i}) := &\, p^* K[p^* - p_i]_+ - \frac{(K[p^* - p_i]_+)^2}{2K} \\ &- C_i\left(K[p^* - p_i]_+\right) \\ \text{s.t.:} \quad K[p^* - p_i]_+ &= D(p^*) - \sum_{j\neq i} K[p^* - p_j]_+\,. \end{aligned} \tag{6}$$

We can now state the following corollary that formalizes what previously observed, which is a direct consequence of Lemma 1.

Corollary 1. The set of Nash equilibria of the PAB auction $\mathcal{U}$ coincides with the set of Nash equilibria of the restricted game $\mathcal{U}^r$.

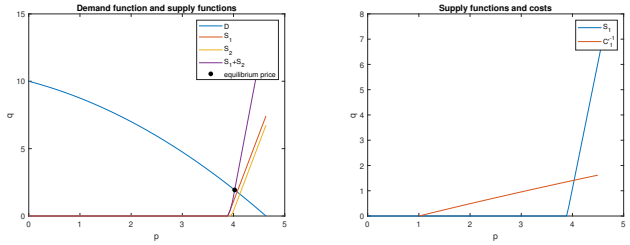


Fig. 3. The unique equilibrium **S** for Example 1

Therefore, it is sufficient to study the restricted game $\mathcal{U}^r$ to determine the entire set of Nash equilibria of the game $\mathcal{U}$. In the next section, we shall prove that the game $\mathcal{U}_r$ admits at least one Nash equilibrium.

3.2 Existence and characterization of Nash equilibria

In this section, we shall prove that the restricted game $\mathcal{U}_r$ admits at least one Nash equilibrium. According to Corollary 1, the Nash equilibrium of the restricted game $\mathcal{U}_r$ corresponds to a Nash equilibrium of the PAB auction $\mathcal{U}$ with strategy space $\mathcal{A}_K$. Therefore, the PAB auction $\mathcal{U}$ with strategy space $\mathcal{A}_K$ admits Nash equilibria that can be characterized in the same way.

Theorem 1. Consider the PAB action with strategy space $\mathcal{A} = \mathcal{A}_K$ as in (4). Then, there exists at least one Nash equilibrium **S** of the form $S_i(p) = K[p - p_i]_+$ for $i \in \{1, 2\}$ and $p_i \in [0, \hat{p}]$.

To prove the statement, we show that the utility function in (6) is quasi-concave in p_i. Quasi-concavity is proved using the fact that the utility is concave in the stationary points. This property, combined with the compactness and convexity of $[0, \hat{p}]$ and the continuity of u_i^r, permits to apply Proposition 20.3 in Osborne and Rubinstein (1994) (pp. 19-20) which guarantees existence of pure-strategy Nash equilibria.

Let us conclude with an example.

Example 1. Let $K = 10$, $D(p) = -\frac{1}{4}p^2 - p + 10$, $C_1(q) = \frac{1}{30}q^3 + q^2 + q$ and $C_2(q) = \frac{1}{30}q^3 + \frac{3}{2}q^2 + 2q$. Then, there exists a unique equilibrium $S = (S_1, S_2)$ where

$$S_1(p) = 10[p - 3.89]_+$$
$$S_2(p) = 10[p - 3.96]_+$$

as shown in Fig. 3

4. CONCLUSION

We considered a supply function equilibrium model with pay-as-bid remuneration and asymmetric firms. Existence of an equilibrium is ensured if we restrict the strategy space to K-Lipschitz supply functions. In such setting, a characterization of Nash equilibria is given.

Further work comprehends conditions for uniqueness of Nash equilibria. Also, we intend to include uncertainty in our model. Motivated by the current structure of electricity markets, we aim to study the concatenation of a uniform-price auction and a pay-as-bid one, modeled as a two-stage game.

REFERENCES

Anderson, E.J. and Philpott, A.B. (2002). Using supply functions for offering generation into an electricity market. *Operations research*, 50(3), 477–489.

Baldick, R., Grant, R., and Kahn, E. (2004). Theory and Application of Linear Supply Function Equilibrium in Electricity Markets. *Journal of Regulatory Economics*, 25(2), 143–167.

Baldick, R. and Hogan, W. (2001). Capacity constrained supply function equilibrium models of electricity markets: Stability, nondecreasing constraints, and function space iterations.

David, A.K. (1993). Competitive bidding in electricity supply. In *IEE proceedings C-Generation, transmission and distribution*, volume 140, 421–426. IET.

Fabra, N., von der Fehr, N.H., and Harbord, D. (2006). Designing electricity auctions. *The RAND Journal of Economics*, 37(1), 23–46.

Genc, T.S. (2009). Discriminatory versus uniform-price electricity auctions with supply function equilibrium. *Journal of optimization theory and applications*, 140(1), 9–31.

Gestore dei Mercati Energetici (GME) (2022). Spot Electricity Market (MPE) - MGP, MI, MPEG, MSD). `https://www.mercatoelettrico.org/en/mercati`.

Green, R. and Newbery, D.M. (1992). Competition in the british electricity spot market. *Journal of Political Economy*, 100(5), 929–53.

Holmberg, P. and Newbery, D. (2010). The supply function equilibrium and its policy implications for wholesale electricity auctions. *Utilities Policy*, 18(4), 209–226.

Holmberg, P. (2009). Supply function equilibria of pay-as-bid auctions. *Journal of Regulatory Economics*, 36, 154–177. doi:10.1007/s11149-009-9091-6.

Hortaçsu, A. and Puller, S.L. (2008). Understanding strategic bidding in multi-unit auctions: a case study of the texas electricity spot market. *The RAND Journal of Economics*, 39(1), 86–114.

Klemperer, P. and Meyer, M. (1989). Supply function equilibria in oligopoly under uncertainty. *Econometrica*, 57(6), 1243–77.

Mas-Colell, A., Whinston, M.D., and Green, J.R. (1995). *Microeconomic Theory*. Number 9780195102680 in OUP Catalogue. Oxford University Press.

Müsgens, F., Ockenfels, A., and Peek, M. (2014). Economics and design of balancing power markets in germany. *International Journal of Electrical Power & Energy Systems*, 55, 392–401.

Osborne, M.J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Rassenti, S.J., Smith, V.L., and Wilson, B.J. (2003). Discriminatory price auctions in electricity markets: low volatility at the expense of high price levels. *Journal of regulatory Economics*, 23(2), 109–123.

Sioshansi, R. and Oren, S. (2007). How good are supply function equilibrium models: an empirical analysis of the ERCOT balancing market. *Journal of Regulatory Economics*, 31(1), 1–35.

Ventosa, M., Baillo, A., Ramos, A., and Rivier, M. (2005). Electricity market modeling trends. *Energy policy*, 33(7), 897–913.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Linear-Matrix-Inequality criteria for various notions of stability/performance ⋆

Joseph A. Ball * **Vladimir Bolotnikov** **

* *Department of Mathematics, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: joball@math.vt.edu).*
** *Department of Mathematics, William & Mary, Williamsburg, VA 23187 USA (e-mail: vladi@math.wm.edu).*

Abstract: We review Linear-Matrix-Inequality criteria for various notions of stability and performance for discrete-time autonomous (state/output) linear systems. Here we discuss analogues of all these ideas for a certain class of time-varying discrete-time, state/output linear systems, where output-stability requires that the Z-transform of the output sequence be in a weighted Bergman space over the unit disk rather than in the usual Hardy space over the unit disk.

Keywords: stability, time-varying system, linear matrix inequality, weighted Bergman space

1. INTRODUCTION

Here $\mathcal{X}$ (the state space) and $\mathcal{Y}$ (the output space) are Hilbert spaces and we suppose that we are given an output pair $(C, A) \in \mathcal{L}(\mathcal{X}, \mathcal{Y}) \times \mathcal{L}(\mathcal{X})$ generating a state/output linear system

$$\Sigma = \Sigma_{C,A}: \begin{cases} x(j+1) &= Ax(j) \\ y(j) &= Cx(j), \end{cases} \quad j \in \mathbf{Z}_+,$$

there are various notions of stability:

(1) Σ is *internally stable*:

$$x(0) \text{ arbitrary } \Rightarrow \lim_{j\to\infty} \|x(j)\| = 0,$$

i.e., A is *strongly stable:* $\lim_{j\to\infty} \|A^j x_0\| = 0$ for each $x_0 \in \mathcal{X}$.

(2) Σ is *internally exponentially stable*, i.e.,

$$\exists\, \rho < 1,\ M < \infty \text{ so that } \|x(j)\| \le M\rho^j \|x(0)\|.$$

Equivalent conditions are:

(2a) A is similar to a strict contraction operator: $\exists$ an invertible X on $\mathcal{X}$ so that $\|X^{-1}AX\| < 1$.

(2b) A has spectral radius strictly less than 1:

$$\rho_{\rm sp}(A) := \lim_{j\to\infty} \|A^j\|^{\frac{1}{j}} < 1.$$

(2c) The inverse linear pencil $L(\lambda)^{-1} = (I - \lambda A)^{-1}$ exists and is uniformly bounded on the closed unit disk.

(3) Σ is *output stable*, i.e.,

$$x(0) \text{ arbitrary} \Rightarrow \{y(j)\}_{j\ge 0} \in \ell^2_{\mathcal{Y}}.$$

We note that the state/output-signal map is given explicitly by the *observability operator* $\mathcal{O}^{\circ}_{C,A}: \operatorname{col}[CA^j]_{j\ge 0}$ being bounded as an operator from $\mathcal{X}$ to $\ell^2_{\mathcal{Y}}$. We prefer to work with the Z-transformed version of the observability operator $\mathcal{O}^{\circ}_{C,A}$ given simply by

$$\mathcal{O}_{C,A}: x \mapsto \sum_{j=0}^{\infty} (CA^j x)\lambda^j.$$

Conditions equivalent to output-stability of the output pair (C, A) are:

(3a) The observability operator $\mathcal{O}_{C,A}$ is bounded as an operator from $\mathcal{X}$ to $H^2_{\mathcal{Y}}$.

(3b) The observability gramian $\mathcal{G}_{C,A} = \mathcal{O}^*_{C,A}\mathcal{O}_{C,A}$ exists as a bounded operator on $\mathcal{X}$, and is given by the strongly convergent infinite series

$$\mathcal{G}_{C,A} = \sum_{j=0}^{\infty} A^{*j} C^* C A^j.$$

While there seems to be no known Linear-Matrix-Inequality (LMI) criterion for the property (1) (except in the finite-dimensional case), there are LMI criteria for the notions (2), (3):

(2) Exponential stability holds if and only if there is a bounded positive definite $H \succ 0$ such that

$$H - A^* H A \succ 0. \tag{1}$$

(3) Output stability holds if and only if there is a bounded positive semi-definite $H \succeq 0$ on $\mathcal{X}$ satisfying the Stein inequality

$$H - A^* H A \succeq C^* C. \tag{2}$$

Specializing to the special case where $\mathcal{X} = \mathcal{Y}$ and $C = I_{\mathcal{X}}$, we arrive at the following statement: *Given $A \in \mathcal{L}(\mathcal{X})$, A has the property that $\sum_{j=0}^{\infty} \|A^j x\|^2 < \infty$ for any $x_0 \in \mathcal{X}$ if and only if $\exists\ H \succeq 0$ so that $H - A^* H A \succeq I_{\mathcal{X}}$.* Since the right-hand side of this last equation is strictly positive definite, we see that its solution H must also be strictly positive definite. Then a rescaling of this latter inequality finally gets us back to (1). We conclude that *exponential stability* for the state operator A is equivalent to output stability for the output pair $(I_{\mathcal{X}}, A)$, a fact presumably which one can also see directly.

⋆ The second author was supported by a collaboration grant from the Simons Foundation.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Whether or not A is strongly stable (the *internal stability* property (1) above) plays a key role in the finer structure of the solution set for inequality (2), as illustrated in the following result.

Theorem 1. Suppose that the output pair (C, A) is output stable, so the inequality (2) has a positive semidefinite solution H. Then:

(1) The observability gramian $H = \mathcal{G}_{C,A}$ is the minimal positive semidefinite solution of (2) and satisfies (2) with equality:
$$H - A^*HA = C^*C. \tag{3}$$
(2) If A is strongly stable, then $H = \mathcal{G}_{C,A}$ is the unique solution of the equality (3).
(3) Suppose that there exists $H' \succ 0$ such that $H' - A^*H'A \succ 0$ (equivalently, A is similar to a strict contraction). Then conversely, if $H = \mathcal{G}_{C,A}$ is the unique solution of (3), then A is strongly stable.

Our goal here is to discuss recent work of the authors (see Ball-Bolotnikov (2013) and the more comprehensive treatment Ball-Bolotnikov (2022)) on a certain class of time-varying discrete-time state/output linear systems. In the definition of output stability, rather than requiring that the Z-transform $\widehat{y}(\lambda) = \sum_{j=0}^{\infty} y_j\lambda^j$ of the output sequence $\{y(j)\}_{j\geq 0}$ be in the Hardy space H^2 (i.e., that $\sum_{j=0}^{\infty}\|y(j)\|^2 < \infty$), we require that $\widehat{y}(\lambda)$ be in a prescribed weighted Bergman space over the unit disk. Our motivation comes from operator theory (the study of the model theory for various classes of hypercontraction operators on a Hilbert space), but related problems with a more engineering motivation have been studied in a continuous-time setting in the work of Partington and collaborators (see e.g. Jacob-et-al (2018)).

2. WEIGHTED BERGMAN SPACES

We note that the Z-transform of the space $\ell^2_{\mathcal{Y}}$ which has a prominent role in the discussion of the previous section is the Hardy space $H^2_{\mathcal{Y}}$

$$H^2_{\mathcal{Y}} = \{f(\lambda) = \sum_{j=0}^{\infty} f_j\lambda^j : \sum_{j=0}^{\infty}\|f_j\|^2_{\mathcal{Y}} < \infty\}.$$

The classical Bergman space $\mathcal{A}_2$ (here taken with values in the coefficient Hilbert space $\mathcal{Y}$) consists of holomorphic $\mathcal{Y}$-valued functions on the unit disk which are square-integrable with respect to area measure:

$$\mathcal{A}_{2,\mathcal{Y}} = \{f(\lambda) = \sum_{j=0}^{\infty} f_j\lambda^j : \int_{\mathbf{D}}\|f(\lambda)\|^2_{\mathcal{Y}}dA(\lambda) < \infty\}$$

where the area integral can also be expressed in terms of Taylor-series coefficients:

$$\|f\|^2_{\mathcal{A}_{2,\mathcal{Y}}} = \sum_{j=0}^{\infty}\frac{1}{j+1}\|f_j\|^2_{\mathcal{Y}}$$

More generally, for any natural number $n \geq 2$ we may consider the space $\mathcal{A}_{n,\mathcal{Y}}$ consisting of $\mathcal{Y}$-valued holomorphic functions on $\mathbf{D}$ having finite weighted area integral

$$\|f\|^2_{\mathcal{A}_{n,\mathcal{Y}}} = \frac{1}{\pi}\int_{\mathbf{D}}\|f(\lambda)\|^2_{\mathcal{Y}}(n-1)(1-|\lambda|^2)^{n-2}dA(\lambda)$$

or in terms of Taylor coefficients

$$\|f\|^2_{\mathcal{A}_{n,\mathcal{Y}}} = \sum_{j=0}^{\infty}\mu_{n,j}\|f_j\|^2_{\mathcal{Y}}$$

where $\mu_{n,j} = \frac{j!}{n(n+1)\cdots(n+j+1)}$ are reciprocal binomial coefficients. All these spaces embed into a family of spaces with continuous index ρ restricted only by $1 < \rho < \infty$ where

$$\|f\|^2_{\mathcal{A}_\rho} = \frac{1}{\pi}\int_{\mathbf{D}}\|f(\lambda)\|^2_{\mathcal{Y}}dA_\rho(\lambda)$$

and $dA_\rho(\lambda) = (\rho-1)(1-|\lambda|^2)^{\rho-2}dA(\lambda)$ or equivalently by

$$\|f\|^2_{\mathcal{A}_\rho} = \sum_{j=0}^{\infty}\mu_{\rho,j}\|f_j\|^2_{\mathcal{Y}}$$

with $\mu_{\rho,j} = \frac{j!}{\rho(\rho+1)\cdots(\rho+j-1)} = \frac{j!\Gamma(\rho)}{\Gamma(\rho+j)}$ where Γ is the usual gamma function meromorphic on the whole complex plane $\mathbf{C}$ with simple poles at the negative integers $-1, -2, \ldots$ which satisfies the interpolation conditions $\Gamma(n) = (n-1)!$. Still more generally, following Ball-Bolotnikov (2013), we introduce *admissible weights* $\omega = \{\omega_j\}_{j\geq 0}$ satisfying admissibility conditions

$$\omega_0 = 1, \quad 1 \leq \frac{\omega_j}{\omega_{j+1}} \leq M \text{ for all } j \text{ for some } M < \infty. \tag{4}$$

A useful consequence of (4) is that the sequence $\{\omega_j^{-1}\}_{j\geq 0}$ is nondecreasing:

$$1 = \omega_0^{-1} \leq \omega_1^{-1} \leq \cdots \leq \omega_j^{-1} \leq \omega_{j+1}^{-1} \leq \cdots. \tag{5}$$

We then define a generalized weighted Bergman space $\mathcal{A}_{\omega,\mathcal{Y}}$ to consist of all $\mathcal{Y}$-valued functions on $\mathbf{D}$ with finite $\mathcal{A}_{\omega,\mathcal{Y}}$-norm, where

$$\|\sum_{j=0}^{\infty} f_j\lambda^j\|^2_{\mathcal{A}_{\omega,\mathcal{Y}}} = \sum_{j=0}^{\infty}\omega_j\|f_j\|^2_{\mathcal{Y}}.$$

One can show that $\mathcal{A}_{\omega,\mathcal{Y}}$ is a reproducing kernel Hilbert space with reproducing kernel of the form

$$K_{\mathcal{A}_{\omega,\mathcal{Y}}}(\lambda,\eta) = k_\omega(\lambda,\eta)I_{\mathcal{Y}}$$

where the scalar-valued kernel k_ω is given by

$$k_\omega(\lambda,\eta) = \sum_{j=0}^{\infty}\omega_j^{-1}\lambda^j\overline{\eta}^j.$$

We assume that $k_\omega(\lambda,\eta) \neq 0$ for λ, η in the unit disk $\mathbf{D}$. We can then expand the inverse function $\frac{1}{k_\omega}$ into a series of the form

$$\frac{1}{k_\omega}(\lambda,\eta) = \sum_{j=0}^{\infty}c_j\lambda^j\overline{\eta}^j.$$

We shall also occasionally need the *shifted kernel function*: for $\kappa = 1, 2, \ldots$, we define

$$k_\omega^{(\kappa)}(\lambda,\eta) = \sum_{j=0}^{\infty}\omega_{j+\kappa}^{-1}\lambda^j\overline{\eta}^j.$$

We shall need the additional hypothesis concerning the coefficients $\{c_j\}_{j\geq 0}$ above:

$$\sum_{j=0}^{\infty}|c_j| < \infty. \tag{6}$$

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Let us also introduce a quick and dirty functional calculus for functions of the form $f(\lambda,\eta)=\sum_{j=0}^{\infty} f_j\lambda^j\overline{\eta}^j$. For S,T,H operators on $\mathcal{X}$, we define

$$f(S,T)[H]=\sum_{j=0}^{\infty} f_j S^j H T^{*j}\in\mathcal{L}(\mathcal{X}) \tag{7}$$

whenever the series converges. We shall apply this functional calculus with f equal to the following functions:

$$f=\frac{1}{k_\omega},\quad f=\frac{k_\omega^{(\kappa)}}{k_\omega}\text{ for }\kappa=1,2,\dots. \tag{8}$$

3. NOTIONS OF WEIGHTED STABILITY

Associated with an output pair $(C,A)\in\mathcal{L}(\mathcal{X},\mathcal{Y})\times\mathcal{L}(\mathcal{X})$ is the weighted state/output linear system

$$\Sigma_\omega=\Sigma_{\omega,C,A}\colon\begin{cases} x(j+1) &= \dfrac{\omega_j}{\omega_{j+1}}Ax(j),\\ y(j) &= Cx(j).\end{cases}$$

Solving the recursions leads to

$$\begin{aligned} x(j) &= \omega_j^{-1}A^jx(0),\\ y(j) &= \omega_j^{-1}CA^jx(0).\end{aligned}$$

There are various possibilities for notions of stability for such an autonomous time-varying system but it is not clear which ones are useful and analyzable. For example one might formulate *internal stability* of Σ_ρ (or ω-strong stability of the operator A) to mean that $\lim_{j\to\infty}x(j)=0$ whenever $\{x(j)\}_{j\ge0}$ is a state trajectory of Σ_ω, or equivalently, *A is ω-strongly stable* in the sense that $\lim_{j\to\infty}\omega_j^{-1}A^jx_0=0$ for all $x_0\in\mathcal{X}$, but we shall introduce below another notion of *ω-strong stability* for A which appears to be more useful. We focus here on:

(2-ω) Σ_ω is *internally ω-exponentially stable*, i.e.,

$$\exists\rho<1,\ M<\infty\text{ so that }\|x(j)\|\le M\rho^j\|x(0)\|,$$

or A is *ω-exponentially stable:* $\exists\rho<1$, $M<\infty$ so that

$$\omega_j^{-1}\|A^jx_0\|\le M\rho^j\|x_0\|\,\forall x_0\in\mathcal{X}.$$

An equivalent condition is:

(2b-ω) A has ω-spectral radius less than 1, i.e.,

$$\rho_{\omega,\mathrm{sp}}(A):=\limsup_{j\to\infty}\left\{\omega_j^{-\frac{1}{j}}\|A^j\|^{\frac{1}{j}}\right\}<1.$$

(3-ω) Σ_ω is *ω-output stable,* i.e.,

$$x(0)\text{ arbitrary }\Rightarrow\sum_{j=0}^{\infty}y(j)\lambda^j\in H^2_\omega.$$

Equivalent conditions are:

(3a-ω) the *ω-observability operator*

$$\begin{aligned}\mathcal{O}_{\omega,C,A}:\ x\mapsto C\big(\sum_{j=0}^{\infty}\omega_j^{-1}\lambda^jA^j\big)x\\ = Ck_\omega(\lambda I_{\mathcal{X}},A^*)x\end{aligned}$$

maps $\mathcal{X}$ boundedly into $H^2_{\omega,\mathcal{Y}}$.

(3b-ω) The weighted observability gramian

$$\mathcal{G}_{\omega,C,A}=\mathcal{O}^*_{\omega,C,A}\mathcal{O}_{\omega,C,A}$$

is a bounded operator on $\mathcal{X}$.

We conjecture: $A\in\mathcal{L}(\mathcal{X})$ is *ω-exponentially stable* if and only if the output pair $(C,I_{\mathcal{X}})$ is ω-output stable.

The LMI criterion for ω-output stability is as follows.

Theorem 2. The output pair $(C,A)\in\mathcal{L}(\mathcal{X},\mathcal{Y})\times\mathcal{L}(\mathcal{X})$ is ω-output stable if and only if there exists $H\in\mathcal{L}(\mathcal{X})$ so that

$$H\succeq A^*HA\succeq0, \tag{9}$$

$$\frac{k_\omega^{(\kappa)}}{k_\omega}(A^*,A)[H]\succeq0\text{ for }\kappa=1,2,\dots,\text{ and} \tag{10}$$

$$\frac{1}{k_\omega}(A^*,A^*)[H]\succeq C^*C. \tag{11}$$

Note that replacing C^*C with $I_{\mathcal{X}}$ gives a criterion for ω-exponential stability for A, assuming the validity of the conjecture just preceding Theorem 2.

The statement of the analogue of Theorem 1 requires a strengthening of the notion of *strong stability* to a notion of *ω-strong stability: Given an operator A on the Hilbert space $\mathcal{X}$, we say that A is* ω-strongly stable *if*

$$\lim_{N\to\infty}A^{*N}\left(\frac{k_\omega^{(N)}}{k_\omega}(A^*,A^*)[I_{\mathcal{X}}]\right)A^N=0$$

in the strong operator topology.

Then we have the following result.

Theorem 3. (See Section 3 in Ball-Bolotnikov (2013).) Suppose that (C,A) is an ω-stable output pair.

(1) Then $H=\mathcal{G}_{\omega,C,A}$ is the minimal positive semidefinite solution of inequality (11) and itself satisfies this inequality with equality:
$$\frac{1}{k_\omega}(A^*,A^*)[H]=C^*C\text{ for }H=\mathcal{G}_{\omega,C,A}. \tag{12}$$
(2) Suppose that $H=I_{\mathcal{X}}$ satisfies all the inequalities (9), (10), (11). Then $\mathcal{G}_{\omega,C,A}\preceq I_{\mathcal{X}}$, i.e., the ω-observability operator $\mathcal{O}_{\omega,C,A}$ is a contraction.
(3) The ω-observability gramian $\mathcal{G}_{\omega,C,A}$ is equal to the identity $I_{\mathcal{X}}$ (i.e., the ω-observability operator $\mathcal{O}_{\omega,C,A}$ is isometric) if and only if (i) (C,A) satisfies inequalities (9), (10) and equality (12), and (ii) A is ω-strongly stable.

Remark. When we specialize the general admissible weight ω to one of the standard weights μ_n $(n=1,2,\dots)$, Theorems 2 and 3 simplify considerably:

- The infinite set of inequalities (10) holds automatically and one is left only with the two inequalities (9), (11).
- μ_n-strong stability is the same as strong stability.
- One can add a uniqueness statement to item (1) in Theorem 3: *for the case where $\omega=\mu_n$ for some $n=2,3,\dots$, $\mathcal{G}_{\mu_n,C,A}$ is the unique solution to (11) if and only if A is strongly stable.*

When (9) and (11) hold with $H=I_{\mathcal{X}}$ and $C=0$, then A is said to be an *n-hypercontraction* in the operator theory literature (see e.g. Agler (1985)). We expect that recent operator-theory work of Olofsson (2015) should adapt to provide a sharpening of Theorems 2 and 3.

REFERENCES

J. Agler, *Hypercontractions and subnormality*, J. Operator Theory **13** (2) (1985), 203-217.

J.A. Ball and V. Bolotnikov, *Weighted Hardy spaces: shift invariant and coinvariant subspaces, linear systems and operator model theory* Acta Sci Math. (Szeged) **79** (2013b), 623-686.

J.A. Ball and V. Bolotnikov, *Noncommutative Function-Theoretic Operator Theory and Applications*, Cambridge Tracts in Mathematics **225**, Cambridge University Press, 2022.

B. Jacob, J.R. Partington, S. Pott, and A. Wynn, *β-admissibility of observation operators for hypercontractive semigroups*, J. Evol. Equ. **18** (2018), 153-170.

A. Olofsson, *Parts of adjoint weighted shifts*, J. Operator Theory **74** (2) (2015), 249-280.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Noncommutative Nullstellensätze and Perfect Games ⋆

Adam Bene Watts * **J. William Helton** ** **Igor Klep** ***

* *University of Waterloo, Canada.*
(e-mail: adam.benewatts1@uwaterloo.ca).
** *University of California San Diego, USA.*
(e-mail: helton@math.ucsd.edu)
*** *University of Ljubljana, Slovenia. (e-mail: igor.klep@fmf.uni-lj.si)*

Abstract: The foundations of classical Algebraic Geometry and Real Algebraic Geometry are the Nullstellensatz and Positivstellensatz. Over the last two decades the basic analogous theorems for matrix and operator theory (noncommutative variables) have emerged. This paper concerns commuting operator strategies for nonlocal games, recalls NC Nullstellensatz which are helpful, extends these, and applies them to a very broad collection of games.
The main results of this procedure are two characterizations, based on Nullstellensatz, which apply to games with perfect commuting operator strategies. The first applies to all games and reduces the question of whether or not a game has a perfect commuting operator strategy to a question involving left ideals and sums of squares. Previously, Paulsen and others translated the study of perfect synchronous games to problems entirely involving a $*$-algebra. The characterization we present is analogous, but works for all games. The second characterization is based on a new Nullstellensatz we derive in this paper. It applies to a class of games we call torically determined games, special cases of which are XOR and linear system games. For these games we show the question of whether or not a game has a perfect commuting operator strategy reduces to instances of the subgroup membership problem.

Keywords: Nonlocal Games, Operator Algebras, Noncommutative Algebraic Geometry

1. EXTENDED ABSTRACT

A nonlocal game describes a test performed between a verifier and k players, in which the verifier tests the players' ability to produce correlated responses without communicating. In a round of the game the verifier sends questions to the players and the players return responses to the verifier. The list of questions and responses is then scored according to a function known by both the verifier and the players before the game began. By convention, the score achieved lies in the interval $[0, 1]$. The players cooperate to try and achieve the highest possible score, with the challenge that the players can't communicate while the game is in progress and so don't know the questions sent to other players.

The optimal score the players can achieve on a nonlocal game $\mathcal{G}$ depends on the resources the players share. If the players share only classical randomness the optimal score they can achieve in expectation is called the the classical value of the game, denoted $\omega(\mathcal{G})$. If players share an arbitrary state in a (possibly infinite dimensional) Hilbert space and can make commuting measurements on it the optimal score they can achieve is called the commuting operator value of the game, denoted $\omega^*_{\rm co}(\mathcal{G})$. The supremum value achievable by players who make commuting measurements on a state in a finite dimensional entangled space is called the quantum value, denoted $\omega^*_{\rm q}(\mathcal{G})$. These three values can all differ, though the inequalities $\omega(\mathcal{G}) \leq \omega^*_{\rm q}(\mathcal{G}) \leq \omega^*_{\rm co}(\mathcal{G})$ are always satisfied.

Starting roughly in this century the classical subject of real algebraic geometry has been extended to matrix and operator (noncommutative) variables. Here inequalities and equalities are explained by being equivalent to algebraic formulas, often involving Sums of Squares (**SOS**). These go under the names of **Positivstellensatz** for inequalities and **Nullstellensatz** for equations. Of course finding quantum strategies for games leads to many such noncommutative (**NC**) inequalities and equalities.

In this paper we describe how the well developed NC real algebraic geometry theory applies and integrates with nonlocal games and commuting operator strategies for them. We show a connection between NC Nullstellensatz and whether or not a nonlocal game has a perfect commuting operator solution (i.e., $\omega^*_{\rm co}(\mathcal{G}) = 1$). This connection gives a new algebraic characterization which applies to all nonlocal games with commuting operator value exactly equal to one. This characterization provides a unified algebraic framework through which several previous results concerning the commuting operator value of nonlocal games can be understood. For a large class of games it also reduces the question of whether or not a game has perfect commuting operator value to an instance of the subgroup member-

⋆ IK was supported by the Slovenian Research Agency grants J1-2453, N1-0217 and P1-0222. ABW was supported by NSF grant CCF-1729369.

ship problem, providing a potential starting point for the investigation of several yet-to-be studied families of games.

For context, Positivstellensätze have long played a major role in the study of nonlocal games in that they are behind the standard Navascués et al. (2008); Doherty et al. (2008) upper bound on the commuting operator value of a game. Underlying this bound is one of the earliest NC Positivstellensätze, Helton and McCullough (2004). This paper turns its attention to developing the analogous NC real algebraic geometry which bears on perfect games.

In the remainder of this abstract we introduce some new terminology, review some previous results concerning the commuting operator value of nonlocal games, and then give formal statements of some of our main results.

1.1 Algebraic Description of Commuting-Operator Strategies

A commuting operator strategy for a nonlocal game is a description of how players can use commuting operator measurements to map questions sent by the verifier to responses. Formally, a (commuting operator) strategy can be specified by a Hilbert space $\mathcal{H}$, a state $\psi \in \mathcal{H}$ which is shared by the players and projectors $\{E(\alpha)_a^i\}$ acting on $\mathcal{H}$, where α ranges over all players, i ranges over all questions, and a ranges over all responses. The projector $E(\alpha)_a^i$ can be read as "the projector corresponding to player α giving response a to question i".

Because the Hilbert space $\mathcal{H}$ on which they act is arbitrary, it is difficult to reason about the $E(\alpha)_a^i$ directly. Instead we introduce the universal game algebra $\mathscr{U}$, a $*$-algebra generated by variables $e(\alpha)_a^i$ which satisfy the same relations as the projectors $E(\alpha)_a^i$, for example, that $e(\alpha)_a^i$ and $e(\beta)_b^j$ commute for any $\alpha \neq \beta$.[1] Commuting operator strategies can then be specified by tuples (π, ψ), consisting of a $*$-representation π mapping $\mathscr{U}$ to bounded operators on a Hilbert space $\mathcal{H}$, along with a state $\psi \in \mathcal{H}$. When specified in this way, it is understood that projectors $E(\alpha)_a^i$ are given by $\pi(e(\alpha)_a^i)$ and that ψ gives the state shared by the players.

1.2 Other Characterizations of Perfect Commuting-Operator Strategies

Several other papers have considered the problem of deciding whether or not a game has a perfect commuting operator strategy and given criteria which determine the existence of perfect commuting operator strategies for specific families of nonlocal games. We review some of those families of games and the associated characterizations below.

- Linear systems games are two player games based around system of m linear equations on n variables. In Cleve et al. (2017) it was shown that deciding existence of a perfect commuting operator strategy for a binary linear systems game was equivalent to solving an instance of the word problem on a group called the solution group of the game.
- XOR games are k player games which, similarly to linear system games, test satisfiability of a system of m binary equations on kn variables. In Watts and Helton (2020) it was shown that deciding the existence of a perfect commuting operator strategy for an XOR game was equivalent to solving an instance of the subgroup membership problem on a group called the game group.
- Synchronous games are two player nonlocal games which include "consistency-checks", where Alice and Bob are sent the same question and win iff they send the same response. Other than these consistency checks, the questions and winning responses involved in a synchronous game are arbitrary. In Paulsen et al. (2016) it was shown that there was a perfect commuting operator strategy for a coloring game iff a $*$-algebra associated with a single player's operators could be represented into a C^*-algebra with a faithful trace. In Helton et al. (2017) and Kim et al. (2018) this was generalized to synchronous games.

1.3 Our Results

The main results of this paper are two theorems giving algebraic characterizations of games with perfect commuting operator strategies.

A key concept introduced on the way to proving these theorems is the notion of a game being determined by a set of elements $\mathcal{F} \subseteq \mathscr{U}$. Formally we say a game $\mathcal{G}$ is determined by a set of elements $\mathcal{F}$ if, for any commuting operator strategy (π, ψ), we have that (π, ψ) is a perfect commuting operator strategy for $\mathcal{G}$ iff $\pi(f)\psi = 0$ for all $f \in \mathcal{F}$. We also note that any game $\mathcal{G}$ is naturally determined by two sets of elements. The first, $\mathcal{N}$, consists of elements corresponding to projectors onto responses which obtain a score less than 1 on questions asked by the verifier, while the second, $\mathcal{Y}$, consists of elements $y-1$ with each element y corresponding to projectors onto responses which obtain a score of exactly 1.

Our first major theorem follows from combining the notion of sets of elements which determine a game with a result in noncommutative algebraic geometry known as a Nullstellensatz. To state the result formally, let $\mathfrak{L}(\mathcal{X})$ denote the left ideal of $\mathscr{U}$ generated by $\mathcal{X}$ for any set of elements $\mathcal{X} \subseteq \mathscr{U}$ and $\mathrm{SOS}_{\mathscr{U}}$ denote sums of squares in the algebra $\mathscr{U}$. Then the following result holds:

Theorem 1. For a nonlocal game $\mathcal{G}$ determined by a set $\mathcal{F} \subseteq \mathscr{U}$ the following are equivalent:

(i) $\mathcal{G}$ has a perfect commuting operator strategy;
(ii) $-1 \notin \mathfrak{L}(\mathcal{F}) + \mathfrak{L}(\mathcal{F})^* + \mathrm{SOS}_{\mathscr{U}}$.

Theorem 1, combined with the natural determining sets $\mathcal{N}$ and $\mathcal{Y}$, gives a fully algebraic characterization of nonlocal games with perfect commuting operator strategies. This characterization is analogous to the characterization of synchronous games given in Helton et al. (2017), but works for all games. For the special case of synchronous games we show that the characterizations of Theorem 1 and Helton et al. (2017) are equivalent.

The second major theorem focuses on a general class of games on which Theorem 1 can be simplified further. A

[1] The $*$-algebra $\mathscr{U}$ is isomorphic to a group algebra and has appeared before in other contexts. For example, in Lupini et al. (2020) an algebra closely related to $\mathscr{U}$ was denoted $\mathcal{A}(X, A)$.

game $\mathcal{G}$ is called a torically determined game if there exists a group G with $\mathscr{U} \cong \mathbb{C}[G]$ and $\mathcal{G}$ is determined by a set of elements

$$\mathcal{F} = \{\beta_i g_i - 1\} \tag{1}$$

with each $\beta_i \in \mathbb{C}$ and $g_i \in G$. We call the elements $\beta_i g_i$ clauses of $\mathcal{F}$, and let $\mathscr{H} = \{\beta_i g_i\}$ be the set of all the clauses of $\mathcal{F}$. We give the following characterization of torically determined games with perfect commuting operator strategies:

Theorem 2. Let $\mathcal{G}$ be a game torically determined by a set of elements $\mathcal{F}$ with clauses $\mathscr{H}$. Then $\mathcal{G}$ has a perfect commuting operator strategy iff the following equivalent criteria are satisfied:

(i) $-1 \notin \mathfrak{L}(\mathcal{F}) + \mathfrak{L}(\mathcal{F})^*$;
(ii) The subgroup H of $\mathscr{U}$ generated by $\mathscr{H} \cup \mathscr{H}^*$ meets $\mathbb{C}$ only in 1.

Condition (i) makes it clear Theorem 2 can be viewed as a version of Theorem 1 for torically determined games which holds without the SOS term. Additionally, condition (ii) reduces the characterization of perfect commuting operator strategies in terms of $*$-algebras given in Theorem 1 to one entirerly in terms of groups. In the paper we show that both linear systems games and XOR games are torically determined games, and that Theorem 2 recovers the algebraic characterizations of these games given in Cleve et al. (2017); Watts and Helton (2020) respectively. We also show that Theorem 2 lets us extend the algebraic characterization of both XOR and binary linear systems games to more general games based on linear equations Mod r for any integer r. [2]

Both Theorems 1 and 2 allow new algorithms for identifying nonlocal games with perfect commuting operator strategies. We will discuss one such algorithm, based on Gröbner bases, and give some sample applications. We note that, unlike the upper bounds coming from the ncSoS hierarchy, these Gröbner bases algorithms can both prove a game has commuting operator value strictly less than 1 and identify some games with commuting operator value exactly equal to 1. [3]

ACKNOWLEDGEMENTS

The authors would like to thank Vern Paulsen and William Slofstra for helpful discussions.

REFERENCES

Cleve, R., Liu, L., and Slofstra, W. (2017). Perfect commuting-operator strategies for linear system games. *J. Math. Phys.*, 58(1), 012202.

Doherty, A.C., Liang, Y.C., Toner, B., and Wehner, S. (2008). The quantum moment problem and bounds on entangled multi-prover games. In *2008 23rd Annual IEEE Conference on Computational Complexity*, 199–210. IEEE.

Goldberg, A. (2021). Synchronous linear constraint system games. *J. Math. Phys.*, 62(3), 032201.

Helton, J.W. and McCullough, S. (2004). A Positivstellensatz for noncommutative polynomials. *Trans. Amer. Math. Soc.*, 356(9), 3721– 3737.

Helton, J.W., Meyer, K.P., Paulsen, V.I., and Satriano, M. (2017). Algebras, synchronous games and chromatic numbers of graphs. *Preprint.* `https://arxiv.org/abs/1703.00960`.

Kim, S.J., Paulsen, V., and Schafhauser, C. (2018). A synchronous game for binary constraint systems. *J. Math. Phys.*, 59(3), 032201.

Lupini, M., Mančinska, L., Paulsen, V.I., Roberson, D.E., Scarpa, G., Severini, S., Todorov, I.G., and Winter, A. (2020). Perfect strategies for non-local games. *Mathematical Physics, Analysis and Geometry*, 23(1), 1–31.

Navascués, M., Pironio, S., and Acín, A. (2008). A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations. *New J. Phys.*, 10(7), 073013.

Paulsen, V.I., Severini, S., Stahlke, D., Todorov, I.G., and Winter, A. (2016). Estimating quantum chromatic numbers. *J. Funct. Anal.*, 270(6), 2188–2222.

Slofstra, W. (2020). Tsirelson's problem and an embedding theorem for groups arising from non-local games. *J. Amer. Math. Soc.*, 33(1), 1–56.

Watts, A.B. and Helton, J.W. (2020). 3XOR games with perfect commuting operator strategies have perfect tensor product strategies and are decidable in polynomial time. *Preprint.* `https://arxiv.org/abs/2010.16290`.

[2] In Cleve et al. (2017) it was already observed that the characterization of binary linear systems games presented could be generalized to any system of equations mod p for any prime p. This generalization is given explicitly in Goldberg (2021).

[3] The question of whether a game has perfect commuting operator value is undecidable Slofstra (2020), meaning these algorithms (or any algorithms!) cannot always identify games with commuting operator value one, but there are many examples where they do.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

L^∞-admissibility for non-strongly-continuous semigroups

Karsten Kruse * **Felix L. Schwenninger** **

* *Institute of Mathematics, Hamburg University of Technology, 21073 Hamburg, Germany (e-mail: karsten.kruse@tuhh.de).*
** *Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands and with the Department of Mathematics, University of Hamburg, Bundesstraße 55, D-20146 Hamburg, Germany (e-mail: f.l.schwenninger@utwente.nl).*

Abstract:
Input-to-state stability is characterised by admissibility for linear systems, governed by strongly continuous semigroups. Yet, in some applications semigroups may fail to be strongly continuous with respect to the norm of the underlying Banach space. Typical examples are given by shift-semigroups and the Gauß–Weierstraß semigroup on spaces of bounded continuous functions as well as dual semigroups. This requires a suitable theory for this general setting within the framework of so-called bi-continuous semigroups, including proper admissibility concepts. Our contribution mainly focuses on non-trivial variants of results from the classical case. For instance, the recently shown fact that the generator of a strongly continuous semigroup is only admissible if it is a bounded operator, fails for bi-continuous semigroups.

Keywords: input-to-state stability, semigroups of operators, bi-continuous semigroup, admissibility, maximal regularity

1. INTRODUCTION

Input-to-state stability (ISS) is a notion well-understood for control systems modelled by ordinary differential equations, or delay equations. Its infinite-dimensional counterpart, more precisely, in the context of dynamics governed by partial differential equations, is less understood, despite intensive research efforts in the recent 15 years. The so-far existing theory as well as a listing of remaining challenges, such as for instance the existence of ISS Lyapunov functions, is nicely summarized in the recent survey Mironchenko and Prieur (2020), also see Dashkovskiy and Mironchenko (2013), Mironchenko and Wirth (2018), and Karafyllis and Krstic (2019). Clearly, difficulties in the infinite-dimensional situation arise due to the many facets of well-posedness for partial differential equations, but also because of the presence of control acting through the boundary in contrast to distributed controls.

Although ISS has proved successful as a concept for large classes of nonlinear systems, already in comparably simple finite-dimensional examples, one cannot expect standard ISS estimates of the form

$$\|x(t)\|_X \leq \beta(\|x(0)\|,t) + \gamma(\|u\|_{\infty,[0,t]}), \qquad t>0, \quad (1)$$

for inputs $u \in L^\infty(0,\infty;U)$ and corresponding states $x(t) \in X$. Here X and U denote Banach spaces, $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$, where $\mathcal{KL}, \mathcal{K}$ refer to standard Lyapunov classes. The mapping $(u, x(0)) \mapsto x(t))$ can be viewed as solution operator for given (boundary) inhomogeneity u and initial values $x(0)$. The notion of integral ISS on the other hand describes estimates of the form

$$\|x(t)\|_X \leq \beta(\|x(0)\|,t) + \gamma\Big(\int_0^t \mu(\|u(s)\|_U)\mathrm{d}s\Big), \ t>0, \ (2)$$

with $\gamma, \mu \in \mathcal{K}$. While these notions are rather trivially seen to be equivalent for finite-dimensional linear systems $\dot{x}(t) = Ax(t) + Bu(t)$, the situation in general infinite dimensions remains a notorious open question.

More precisely, if A generates a strongly continuous semigroup $(T(t))_{t\geq 0}$ on the Banach space X, and $B \in \mathcal{L}(U, X_{-1})$, with X_{-1} being the completion of X with respect to $(\lambda I - A)^{-1}$ for some λ in the resolvent set $\rho(A)$ of A, then we ask whether the (mild) solution

$$x(t) = T(t)x_0 + \int_0^t T_{-1}(t-s)Bu(s)\mathrm{d}s, \qquad (3)$$

to the problem

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0, \qquad (4)$$

with $x_0 \in X$, satisfies the ISS estimate (1) or (2), respectively. Here $T_{-1}(t)$ denotes the extension of $T(t)$ to the space X_{-1} which exists uniquely as a bounded operator. Note that here the notion of (integral) ISS in particular includes the property that the solution $x(t)$ lies in the space X, which is non-trivial as a-priori $x(t)$ only lies in X_{-1}, which is also the space on which (4) has to be understood. In fact, if $x(t)$ lies in X for all given input functions from the space $L^\infty(0,\infty;U)$, inequality (1) follows automatically. On the other hand, it is not hard to see that integral ISS always implies ISS for linear systems of the above form, which leaves us with the open question

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Does ISS of a linear system imply integral ISS?

Although several situations, under additional assumptions on the strongly continuous semigroup and the spaces X and U, see e.g. Jacob et al. (2018, 2019), are known, this question in its full generality remains open until today. Recently, in Jacob et al. (2022), an example was provided showing the answer is no, if only continuous input functions are considered. This example is pathologic in the way that $B = A_{-1}$, which can be interpreted as the "worst input operator" one may choose. Still, the question whether L^∞-ISS implies L^∞-integral ISS remains untouched by the example. In the same paper, a relation between ISS (or admissibility respectively) and maximal regularity for abstract evolution equations was revealed. More precisely, it was shown that ISS of $B = A_{-1}$ is equivalent to maximal regularity of A, if both notions are considered with respect to the continuous X-valued functions.

This contribution aims to pave the way for a corresponding theory if the strong continuity of the semigroup is dropped. Moreover, we want to study the above central question on ISS for this more general class as well as the relation to maximal regularity with respect to continuous and essentially bounded functions. In particular, we want to explain why the definitions of ISS respectively admissibility and maximal regularity depend on the considered function classes in a subtle way.

1.1 A motivating example

A first attempt to answer the above mentioned question on ISS in the negative is given by the following example, which is standard in semigroup theory.

Consider the space $X = \ell^\infty$ of bounded, complex-valued sequences $(a_n)_{n\in\mathbb{N}}$ with the supremum norm $\|\cdot\|_\infty$. The operators defined by

$$T(t)(x_n)_{n\in\mathbb{N}} := (\mathrm{e}^{-nt}x_n)_{n\in\mathbb{N}}, \qquad t \geq 0,$$

clearly define a semigroup of bounded linear operators on X, which fails to be strongly continuous. As the dual semigroup of a strongly continuous semigroup on ℓ^1, it is clear that $(T(t))_{t\geq 0}$ is bi-continuous with respect to the weak* topology with generator A given by

$$D(A) = \{(x_n)_{n\in\mathbb{N}} \in \ell^\infty \colon (nx_n)_{n\in\mathbb{N}} \in \ell^\infty\},$$
$$A(x_n)_{n\in\mathbb{N}} = (-nx_n)_{n\in\mathbb{N}}.$$

It is not hard to see that the following facts hold.

- $B = A_{-1}$ is ISS, but
- $B = A_{-1}$ is not integral ISS

with respect to inputs u in $L^\infty(0,\infty;X)$. Note that the corresponding statement changes significantly if we replace the space ℓ^∞ by c_0; then $B = A_{-1}$ is not ISS. This is a consequence of a recent result Jacob et al. (2022) stating that L^∞-ISS for $B = A_{-1}$ for a strongly continuous semigroup already implies the boundedness of the generator.

On the first glance, this result seems to resolve the question posed in the introduction completely. A closer look, however, reveals that the considered function space $L^\infty(0,\infty;X)$ does not match the natural setting we encounter for bi-continuous semigroups.

2. NOTIONS AND PRELIMINARIES

We call a triple $(X, \|\cdot\|, \tau)$ a sequentially complete Saks space if $(X, \|\cdot\|)$ is a Banach space, τ is a coarser Hausdorff locally convex topology than the $\|\cdot\|$-topology such that $(X,\tau)'$ is norming, and if $\|\cdot\|$-bounded τ-Cauchy sequences are convergent. In our motivating example $(\ell^\infty, \|\cdot\|_\infty, \sigma(\ell^\infty,\ell^1))$ is the sequentially complete Saks space. Bi-continuous semigroups on a sequentially complete Saks space $(X, \|\cdot\|, \tau)$ were introduced by Kühnemund (2001, 2003) as exponentially bounded semigroups in the space of bounded linear operators $\mathcal{L}(X)$ on X which are (only) τ-strongly continuous and locally sequentially τ-equicontinuous on $\|\cdot\|$-bounded sets. The precise definition looks as follows.

Definition 1. Let $(X, \|\cdot\|, \tau)$ be a sequentially complete Saks space. A family $(T(t))_{t\geq 0}$ in $\mathcal{L}(X)$ is called *τ-bi-continuous semigroup* if

(i) $(T(t))_{t\geq 0}$ is a *semigroup*, i.e. $T(t+s) = T(t)T(s)$ and $T(0) = \mathrm{id}$ for all $t, s \geq 0$,
(ii) $(T(t))_{t\geq 0}$ is *τ-strongly continuous*, i.e. the map $T_x \colon [0,\infty) \to (X,\tau)$, $T_x(t) := T(t)x$, is continuous for all $x \in X$,
(iii) $(T(t))_{t\geq 0}$ is *exponentially bounded*, i.e. there exist $M \geq 1$, $\omega \in \mathbb{R}$ such that $\|T(t)\| \leq Me^{\omega t}$ for all $t \geq 0$,
(iv) $(T(t))_{t\geq 0}$ is *locally bi-equicontinuous*, i.e. for every sequence $(x_n)_{n\in\mathbb{N}}$ in X, $x \in X$ with $\sup_{n\in\mathbb{N}} \|x_n\| < \infty$ and $\tau\text{-}\lim_{n\to\infty} x_n = x$ it holds that

$$\tau\text{-}\lim_{n\to\infty} T(t)(x_n - x) = 0$$

locally uniformly for all $t \in [0,\infty)$.

In order to interpret the (mild) solution (3) there are some adaptations needed in the bi-continuous setting. The generator of a τ-bi-continuous semigroup is defined similar to the one of a $\|\cdot\|$-strongly continuous semigroup, namely, by

$$Ax := \tau\text{-}\lim_{t\to 0^+} \frac{T(t)x - x}{t},\ x \in D(A),$$

where the domain $D(A)$ consists of all $x \in X$ such that this τ-limit exists and $\sup_{t\in(0,1]} \frac{\|T(t)x-x\|}{t} < \infty$. The notion of the extrapolation space X_{-1} and the extrapolated semigroup $(T_{-1}(t))_{t\geq 0}$ was transferred by Budde and Farkas (2019) to the bi-continuous setting and the semigroup $(T_{-1}(t))_{t\geq 0}$ becomes a bi-continuous semigroup with generator A_{-1} on the sequentially complete extrapolated Saks space $(X_{-1}, \|\cdot\|_{-1}, \tau_{-1})$. The integral appearing in (3) is in general not a Bochner integral (in X_{-1}) anymore due to the extrapolated semigroup being only τ_{-1}-strongly continuous. However, one may regard this integral as a Pettis integral in X_{-1} w.r.t. the τ_{-1}-topology under the constraint that $U = X$, $B \in \mathcal{L}(X; X_{-1})$ is in addition τ-τ_{-1}-continuous and the inputs u belong to the space $C_{\tau,b}([0,t];X)$ of τ-continuous $\|\cdot\|$-bounded functions from $[0,t]$ to X. Now, the operator B is called *$C_{\tau,b}$-admissible* for $t > 0$ if this τ_{-1}-Pettis integral belongs to X, and it turns out that the $C_{\tau,b}$-admissibility of $B = A_{-1}$ is equivalent to

$$\int_0^r T(r-s)f(s)\mathrm{d}s \in D(A)$$

for all $f \in C_{\tau,b}([0,t];X)$ and all $r \in [0,t]$. If B is $C_{\tau,b}$-admissible for all $t > 0$, then we call B $C_{\tau,b}$-admissible.

Looking at the space $C_{\tau,b}([0,t];X)$ and aiming for a corresponding result to Jacob et al. (2022) in the bi-continuous framework, one realizes that the definition of L^∞-admissibility needs an adaptation as well. Namely, due to the functions in $C_{\tau,b}([0,t];X)$ being only τ-continuous they may not be $\|\cdot\|$-strongly measurable and therefore it is not guaranteed that the inclusion $C_{\tau,b}([0,t];X) \subset L^\infty(0,t;X)$ holds. Thus one has to replace the $\|\cdot\|$-strong measurability by τ-strong measurability and so the space $L^\infty(0,t;X)$ by

$$L^\infty_\tau(0,t;X) := \mathscr{L}^\infty_\tau(0,t;X)/\mathscr{N}.$$

where $\mathscr{L}^\infty_\tau(0,t;X)$ is the space of functions $f\colon [0,t] \to X$ which are τ-strongly measurable and essentially $\|\cdot\|$-bounded, and $\mathscr{N}$ is the space of functions in $\mathscr{L}^\infty_\tau(0,t;X)$ which vanish Lebesgue almost everywhere. If τ coincides with the $\|\cdot\|$-topology, then this space coincides with $L^\infty(0,t;X)$.

Definition 2. Let $(X,\|\cdot\|,\tau)$ be a sequentially complete Saks space and $(T(t))_{t\geq 0}$ a τ-bi-continuous semigroup on X. Let $t_0 > 0$ and $F(0,t_0;X)$ be a space of functions on $[0,t_0]$ with values in X such that the convolution

$$(T * f)(t) := \int_0^t T(t-s)f(s)\,\mathrm{d}s$$

is a well-defined τ-Pettis integral in X for any $t \in [0,t_0]$. We say that $(T(t))_{t\geq 0}$ satisfies *F-maximal regularity for* t_0 if for all $f \in F(0,t_0;X)$ it holds that $(T*f)(t) \in D(A)$ for all $t \in [0,t_0]$ and $A(T*f) \in F(0,t_0;X)$. We say that $(T(t))_{t\geq 0}$ satisfies *F-maximal regularity* if it satisfies F-maximal regularity for all $t_0 > 0$.

3. RESULTS

3.1 A theorem on maximal regularity w.r.t. $\|\cdot\|_\infty$-norms

It is well-known that maximal regularity with respect to continuous functions is rare for strongly continuous semigroup generators. This is due to Baillon's result given in Baillon (1980), which even holds if the considered semigroup is not strongly continuous.

Theorem 1. (Baillon's result on maximal regularity). Let A be the generator of a analytic semigroup on a Banach space X such that A satisfies $C([0,t];X)$-maximal regularity. If A is unbounded, then X contains an isomorphic copy of c_0.

The phenomenon observed in Section 1.1, however, is not accidental, as the following result shows.

Theorem 2. Let A generate a strongly continuous semigroup on a Banach space X and suppose that the adjoint A' satisfies $C([0,t];X)$-maximal regularity. Then X' contains ℓ^∞.

However, as mentioned above, maximal regularity and admissibility with respect to functions in $C([0,t];X)$ or $L^\infty([0,t];X)$ is not the natural setting as the continuity/measurability relates to the wrong topology. Instead, the spaces $C_{\tau,b}([0,t];X)$ and $L^\infty_\tau(0,t;X)$ should be used.

The ultimate goal of these considerations is to approach a statement of the following form, which is inspired by the corresponding result for strongly continuous semigroups, Jacob et al. (2022).

Conjecture 3. Let $(X,\|\cdot\|,\tau)$ be a sequentially complete Saks space. and $(T(t))_{t\geq 0}$ a τ-bi-continuous semigroup on X with generator $(A, D(A))$ such that $0 \in \rho(A)$. Then the following assertions are equivalent:

(a) A_{-1} is L^∞_τ-admissible.
(b) $\mathrm{Fav}(T) = D(A)$ and $(T(t))_{t\geq 0}$ satisfies the $C_{\tau,b}$-maximal regularity.
(c) $D(A) = X$ and $A\colon (X,\gamma) \to (X,\gamma)$ is continuous.

Here, L^∞_τ-admissibility is defined analogously to $C_{\tau,b}$-admissibility, $\mathrm{Fav}(T)$ denotes the Favard space of T and $\gamma = \gamma(\|\cdot\|,\tau)$ the mixed topology introduced by Wiweger (1961), see Cooper (1978) as well. If τ coincides with the $\|\cdot\|$-topology, then γ coincides with the $\|\cdot\|$-topology too. So, if true, then this conjecture generalises the mentioned result from Jacob et al. (2022). In order to achieve this, we have developed a corresponding theory of sun dual spaces for bi-continuous semigroups in Kruse and Schwenninger (2022); a tool that was pivotal in the norm-strongly continuous case. These results, which are of interest in their own right, extend works by van Neerven (1992).

REFERENCES

Baillon, J.B. (1980). Caractère borné de certains générateurs de semi-groupes linéaires dans les espaces de Banach. *C. R. Acad. Sci. Paris Sér. A-B*, 290(16), A757–A760.

Budde, C. and Farkas, B. (2019). Intermediate and extrapolated spaces for bi-continuous operator semigroups. *J. Evol. Equ.*, 19(2), 321–359. doi:10.1007/s00028-018-0477-8.

Cooper, J. (1978). *Saks spaces and applications to functional analysis*. North-Holland Math. Stud. 28. North-Holland, Amsterdam.

Dashkovskiy, S. and Mironchenko, A. (2013). Input-to-state stability of infinite-dimensional control systems. *Math. Control Signals Systems*, 25(1), 1–35. doi:10.1007/s00498-012-0090-2.

Jacob, B., Nabiullin, R., Partington, J.R., and Schwenninger, F.L. (2018). Infinite-dimensional input-to-state stability and Orlicz spaces. *SIAM J. Control Optim.*, 56(2), 868–889. doi:10.1137/16M1099467.

Jacob, B., Schwenninger, F.L., and Wintermayr, J. (2022). A refinement of Baillon's theorem on maximal regularity. *Studia Math.*, 263(2), 141–158. doi:10.4064/sm200731-20-3.

Jacob, B., Schwenninger, F.L., and Zwart, H. (2019). On continuity of solutions for parabolic control systems and input-to-state stability. *J. Differential Equations*, 266(10), 6284–6306. doi:10.1016/j.jde.2018.11.004.

Karafyllis, I. and Krstic, M. (2019). *Input-to-state stability for PDEs*. Communications and Control Engineering Series. Springer, Cham. doi:10.1007/978-3-319-91011-6.

Kruse, K. and Schwenninger, F.L. (2022). Sun dual theory for bi-continuous semigroups. ArXiv preprint `https://arxiv.org/abs/2203.12765`.

Kühnemund, F. (2001). *Bi-continuous semigroups on spaces with two topologies: Theory and applications*. Ph.D. thesis, Eberhard-Karls-Universität Tübingen.

Kühnemund, F. (2003). A Hille–Yosida theorem for bi-continuous semigroups. *Semigroup Forum*, 67(2), 205–225. doi:10.1007/s00233-002-5000-3.

Mironchenko, A. and Prieur, C. (2020). Input-to-state stability of infinite-dimensional systems: recent results and open questions. *SIAM Rev.*, 62(3), 529–614. doi:10.1137/19M1291248.

Mironchenko, A. and Wirth, F. (2018). Characterizations of input-to-state stability for infinite-dimensional systems. *IEEE Trans. Automat. Control*, 63(6), 1602–1617. doi:10.1109/tac.2017.2756341.

van Neerven, J. (1992). *The adjoint of a semigroup of linear operators*. Lecture Notes in Math. 1529. Springer, Berlin. doi:10.1007/BFb0085008.

Wiweger, A. (1961). Linear spaces with mixed topology. *Studia Math.*, 20(1), 47–68. doi:10.4064/sm-20-1-47-68.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A non-linear internal model principle for observers ⋆

Jochen Trumpf * **Johannes Nüssle** **

* *Australian National University (e-mail: Jochen.Trumpf@anu.edu.au)*
** *e-mail: nuessle@johannesnuessle.de*

Keywords: observer theory, internal model principle, non-linear observers, behaviours, non-linear systems theory

1. EXTENDED ABSTRACT

State observer design for non-linear systems is concerned with the question of how to construct a dynamical system, the *observer*, that takes as input both the input u and the output y of a given non-linear control system

$$\begin{aligned} \dot{x} &= f(x,u,t), \\ y &= h(x,u,t), \end{aligned} \tag{1}$$

and produces as its output an estimate $\hat{x}$ of the state variable x. The notation in (1) is deliberately generic since many variations of this problem are studied in the literature that differ in terms of the spaces that the variables x, u, and y live in, the assumed properties of the functions f and h, or the notion of what constitutes a good estimate $\hat{x}$ of x.

One way to approach this problem is to make the Ansatz

$$\dot{\hat{x}} = f(\hat{x},u,t) + \Delta(\hat{x},u,y,t), \tag{2}$$

where the *internal model term* $f(\hat{x},u,t)$ simulates the observed system, or *plant*, (1) and the *correction term* $\Delta(\hat{x},u,y,t)$ is zero along trajectories of (1). In the linear context this idea dates back to Luenberger (1964), and in the non-linear context at least to Kou (1973), see also Thau (1973). In the case where $x, \hat{x} \in \mathbb{R}^n$ and where we are seeking an asymptotic observer, the problem is now to construct a correction term $\Delta(\hat{x},u,y,t)$ such that the *observer error* $e := \hat{x} - x$ fulfils

$$\lim_{t\to\infty} e(t) = 0 \tag{3}$$

for all (admissible) choices of $x(0)$, $\hat{x}(0)$ and u.

The significance of $\Delta(\hat{x},u,y,t)$ being zero along trajectories of (1) is that, together with uniqueness of solutions, it implies the *tracking property*

$$\hat{x}(0) = x(0) \implies \hat{x}(t) = x(t) \text{ for all } t. \tag{4}$$

In other words, the behaviour (set of trajectories) of the observer (2) contains the behaviour (set of trajectories) of the plant (1), i.e. an *internal model* of the plant.

An obvious question now is to what extent such an internal model plus correction term design is *necessary*. If we have a general asymptotic observer

$$\dot{\hat{x}} = \hat{f}(\hat{x},u,y,t) \tag{5}$$

for which (3) holds, does the right hand side $\hat{f}$ *always* split as in (2)? An affirmative answer to such a question is called an *internal model principle for observers*. Note that such a result depends on the classes of plants and observers under consideration as well as on the notion of what constitutes a good estimate.

For linear systems, several such internal model principles were proved in Trumpf et al. (2014), covering the most common notions of good estimate: asymptotic, dead-beat, and exact observers. See also the even more general results in Blumthaler and Trumpf (2014). In the linear case, observers do not necessarily contain *full* internal models of the plant but internal models of significant parts of the plant behaviour. For the details see Trumpf et al. (2014) or Blumthaler and Trumpf (2014).

In this work, we will present a general internal model principle for observers formulated in a purely set theoretic generalisation of behavioural observer theory. We show that the historic focus on the linear case has somewhat obscured what is in essence a surprisingly simple theory, at least once the manifold implications of linearity have been disentangled and only the strictly necessary components kept and generalized. We recover the known linear results as special cases of our general result and also derive a novel internal model principle for non-linear kinematic systems on differentiable manifolds. To our best knowledge, this is the first non-linear internal model principle for observers in the literature.

Our theory proceeds from the observations that (3) defines an equivalence relation on the set of state trajectories, i.e. a notion of which pairs of trajectories are *close* to each other, and that sets of trajectories form a *poset* (partially ordered set) under set inclusion. We define the *saturation* of a given behaviour (set of trajectories) as the set closure under the closeness relation and use this concept to define what we call the *radical set* associated with a given behaviour (set of trajectories). The space of saturations of behaviours is a poset under set inclusion. We show that the space of radical sets also carries a natural poset structure. The case where the poset of saturations and the poset of radical sets are isomorphic via the natural isomorphism is of particular interest. We say that the poset of radical sets admits *local poset sections* in this case. This property holds for linear systems as well as for kinematic systems on differentiable manifolds.

⋆ This research is supported by the Australian Research Council Discovery Project DP190103615: "Control of Network Systems with Signed Dynamical Interconnections".

The general internal model principle for observers now states that if the poset of radical sets admits local poset sections and if the radical sets of behaviours are well-founded (the latter is a standard property in the theory of posets that implies the existence of minimal elements) then any non-intrusive, observable observer behaviour contains a minimal element of the radical set of the plant behaviour.

Focusing on the case of asymptotic observer design, for linear systems the minimal element of the radical set is unique and equals the anti-stabilizable part of the plant behaviour, cf. (Trumpf et al., 2014, Theorem 5.6). We show that for kinematic systems on differentiable manifolds the radical set only contains the plant behaviour. It follows that any asymptotic observer for a non-linear kinematic system on a differentiable manifold contains a full internal model of the plant.

REFERENCES

Blumthaler, I. and Trumpf, J. (2014). A new parametrization of linear observers. *IEEE Transactions on Automatic Control*, 59, 1778–1788.

Kou, S. (1973). *Observability and observers for nonlinear dynamic systems*. Ph.D. thesis, Washington University.

Luenberger, D. (1964). Observing the state of a linear system. *IEEE Transactions on Military Electronics*, 8(2), 74–80.

Thau, F. (1973). Observing the state of non-linear dynamic systems. *International Journal of Control*, 17(3), 471–479.

Trumpf, J., Trentelman, H., and Willems, J. (2014). Internal model principles for observers. *IEEE Transactions on Automatic Control*, 59, 1737–1749.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Tropical numerical methods for solving stochastic control problems

Marianne Akian * **Jean-Philippe Chancelier** ** **Luz Pascal** ****
Benoît Tran ***

* *Inria and CMAP, École polytechnique CNRS IP Paris, France (e-mail: marianne.akian@inria.fr)*
** *CERMICS, École des Ponts ParisTech, France (jean-philippe.chancelier@enpc.fr)*
*** *FGV EMAp, Brazil (benoit.tran@tutanota.com)*
**** *Queensland University of Technology & CSIRO, Australia (luz.pascal96@gmail.com)*

Abstract: We consider Dynamic programming equations associated to discrete time stochastic control problems with continuous state space, which arise in particular from monotone time discretizations of Hamilton-Jacobi-Bellman equations. We develop and study several numerical algorithms for solving such equations, combining tropical numerical methods and stochastic dual dynamic programming methods. We also compare these algorithms with the point based methods for solving Partially Observable Markov Decision Processes (POMDP).

Keywords: Stochastic Control, Hamilton-Jacobi-Bellman equations, Stochastic Dual Dynamic Programming, Tropical algebra, Partially Observable Markov decision processes.

1. INTRODUCTION

We consider the following stochastic control problem with discrete time and a possibly discounted additive payoff, either with a finite or infinite horizon T. At each step $t \in [\![0,T]\!]$, the state $\mathbf{X_t} \in \mathbb{X} \subset \mathbb{R}^n$ follows the following dynamics

$$\mathbf{X_{t+1}} = f_t^{\mathbf{W_{t+1}}}(\mathbf{X_t}, \mathbf{U_t}) \ ,$$

where $(\mathbf{W_t})_{t\in[\![0,T]\!]}$ is a sequence of random variables with values in some measurable set $(\mathbb{W}, \mathcal{W})$, and $(\mathbf{U_t})_{t\in[\![0,T]\!]}$ is an *adapted* sequence of (random) decisions or controls with values in some measurable set $(\mathbb{U}, \mathcal{U})$. The state is fully observed, and we may be in the hazard-decision framework in which *adapted* means that, for all t, $\sigma(\mathbf{U_t}) \subset \sigma(\mathbf{X_0}, \mathbf{W_1}, \ldots, \mathbf{W_{t+1}})$. We may also be in the decision-hazard framework, in which *adapted* means that, for all t, $\sigma(\mathbf{U_t}) \subset \sigma(\mathbf{X_0}, \mathbf{W_1}, \ldots, \mathbf{W_t})$. At each time t, the decision maker is receiving the reward

$$r_t^{\mathbf{W_{t+1}}}(\mathbf{X_t}, \mathbf{U_t}) \ ,$$

and at the final time, if any, the decision maker receives the final reward $\psi(\mathbf{X_T})$. Then, the decision maker aims to maximize his total expected reward:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} r_t^{\mathbf{W_{t+1}}}(\mathbf{X_t}, \mathbf{U_t}) + \psi(\mathbf{X_T})\right] \ .$$

Such a problem is also called a multi-stage optimization problem. In the sequel, we shall assume that the random variables $\mathbf{W_{t+1}}$ are independent and with finite support $\mathbb{W}$. The law of $\mathbf{W_{t+1}}$ may depend on t. In the decision-hazard framework, we may also consider the case where the law of $\mathbf{W_{t+1}}$ depends on $(\mathbf{X_t}, \mathbf{U_t})$, which is equivalent to consider the general framework of Markov decision processes, with some given transition probabilities $T_t^{u_t}(x_t, x_{t+1}) = P(\mathbf{X_{t+1}} = x_{t+1} \mid \mathbf{X_t} = x_t, \mathbf{U_t} = u_t)$, such that $T_t^{u_t}(x_t, \cdot)$ has a finite support. In the hazard-decision framework, we can assume that the law of $\mathbf{W_{t+1}}$ depends on $\mathbf{X_t}$, but it cannot depend on $\mathbf{U_t}$. The discrete probability law of $\mathbf{W_{t+1}}$ will be denoted by $p_t^{x_t,u_t}(w) = P(\mathbf{W_{t+1}} = w \mid \mathbf{X_t} = x_t, \mathbf{U_t} = u_t)$ in the first case, and by $p_t^{x_t}(w) = P(\mathbf{W_{t+1}} = w \mid \mathbf{X_t} = x_t)$ in the second case.

By the dynamic programming approach (see Bellman (1984)), the value function of the above problem is the function V_0 obtained from the solution to the following recurrence equation:

$$V_T = \psi \quad \text{and} \quad \forall t \in [\![0, T-1]\!], V_t = \mathfrak{B}_t\left(V_{t+1}\right) \ , \qquad (1)$$

where $\mathfrak{B}_t$ is the associated Bellman operator from the set of extended real functions over $\mathbb{X}$ ($\overline{\mathbb{R}}^{\mathbb{X}}$), to itself. This operator can be written as the composition of three different operators among the following ones which are operating on functions from either $\mathbb{X}$, $\mathbb{X} \times \mathbb{U}$, $\mathbb{X} \times \mathbb{U} \times \mathbb{W}$ or $\mathbb{X} \times \mathbb{W}$ to $\overline{\mathbb{R}}$:

⋆ Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lowercase letters, not all uppercase.

$$\begin{aligned}
\mathcal{Q}_t(\phi)(x,u,w) &= r_t^w(x,u) + \phi\big(f_t^w(x,u)\big),\\
\mathcal{M}_t^{(1)}(Q)(x,w) &= \max_{u\in\mathbb{U}} Q(x,u,w),\\
\mathcal{E}_t^{(2)}(Q)(x) &= \mathbb{E}\Big[Q(x,\mathbf{W_{t+1}})\Big] = \sum_{w\in\mathbb{W}} p_t^x(w)Q(x,w),\\
\mathcal{E}_t^{(1)}(Q)(x,u) &= \mathbb{E}\Big[Q(x,u,\mathbf{W_{t+1}}) \mid \mathbf{X}_t = x, \mathbf{U}_t = u\Big]\\
&= \sum_{w\in\mathbb{W}} p_t^{x,u}(w)Q(x,u,w),\\
\mathcal{M}_t^{(2)}(Q)(x) &= \max_{u\in\mathbb{U}} Q(x,u)\ ,
\end{aligned}$$

in which we use the convention $+\infty-\infty=-\infty$. Indeed, in the hazard-decision case, we have $\mathfrak{B}_t = \mathcal{E}_t^{(2)}\circ\mathcal{M}_t^{(1)}\circ\mathcal{Q}_t$ and in the decision-hazard case, we have $\mathfrak{B}_t = \mathcal{M}_t^{(2)}\circ\mathcal{E}_t^{(1)}\circ\mathcal{Q}_t$.

The above discrete time Bellman equation (1) can also be obtained after some semi-Lagrangian time discretization of a Hamilton-Jacobi-Bellman equation, see Falcone and Ferretti (2014), or any monotone time discretization. The dynamic programming approach suffers from the "curse of dimensionality", since one would need to compute for all $t \in [\![0,T]\!]$ the value function V_t on all the state space $\mathbb{X}$, and any grid-based discretization would need a number of values exponential in the dimension n of the state space $\mathbb{X}$. Several methods have been proposed in the litterature to bypass the obstruction of curse of dimensionality. We shall only cite the ones related to the present work: the tropical numerical methods developped in the context of Hamilton-Jacobi equations (McEneaney (2007); McEneaney et al. (2011); Qu (2014); Akian and Fodjo (2018)), the tree-structured algorithm developped recently by Alla et al. (2019), and the stochastic dual dynamic programming method developped in the context of discrete time stochastic control (Pereira and Pinto (1991); Philpott et al. (2013)).

Here, we consider a general algorithm inspired by both tropical numerical methods and SDDP algorithm and which can be seen as a generalization of the algorithms proposed in Philpott et al. (2013); Baucke et al. (2018)). Moreover, we show that in the case of the dynamic programming equation associated to a partially observable Markov Decision Process (POMDP), it is similar to the so called point based algorithms developped in Pineau et al. (2003); Kurniawati et al. (2008); Shani et al. (2013).

2. TROPICAL NUMERICAL METHOD FOR LIPSCHITZ PRESERVING BELLMAN OPERATORS

The following algorithm is introduced in more details in Akian et al. (2020).

We assume that the map r_t^w takes its values in $\mathbb{R}\cup\{-\infty\}$, to handle constraints in state and control. We also assume that r_t^w is bounded from above, which will imply that the Bellman operator preserves the set of upper bounded function from $\mathbb{X}$ to $\mathbb{R}\cup\{-\infty\}$. For any function ϕ from a set Y to $\mathbb{R}\cup\{-\infty\}$, the support of ϕ will be the set of $y\in Y$ such that $\phi(y)\in\mathbb{R}$. The constraints determined by the support of r_t^w and the support of the final reward ψ induce a sequence of sets $X_t\subset\mathbb{X}$, $t\in[\![0,T]\!]$, such that any sequence V_t, $t\in[\![0,T]\!]$, satisfying (1) is such that the support of V_t is included in X_t.

It is well known that Bellman operators are order preserving and nonexpansive for the sup-norm. Since $\mathbb{X}$ is an infinite set, we shall need the following stronger property:

Assumption 1. There exists a sequence $\mathcal{L}_t$, $t\in[\![0,T]\!]$, of compact subsets of the set of functions from $\mathbb{X}$ to $\mathbb{R}\cup\{-\infty\}$, endowed with the uniform convergence topology, such that, for all $t\in[\![0,T-1]\!]$, the Bellman operator $\mathfrak{B}_t$ sends $\mathcal{L}_{t+1}$ into $\mathcal{L}_t$.

Compact subsets $\mathcal{L}_t$ can be obtained by taking the set of functions from $\mathbb{X}$ to $\mathbb{R}\cup\{-\infty\}$, that are L_t-Lipschitz continuous (for some given norm of $\mathbb{R}^n$) on their compact support X_t, or any closed subset of this set. In Akian et al. (2020), Assumption 1 is proved to be satisfied for these particular sets $\mathcal{L}_t$, for some constants L_t, under some technical conditions similar to the ones that are generally assumed in the proofs of convergence of SDDP algorithm. Another important property assumed to apply SDDP algorithm (in the above context of a maximization problem) is that the control set is polyhedral, that the reward functions are polyhedral and concave with respect to state and control, and that the dynamics are affine with respect to state and control. In that case the value function can be approximated by the finite infimum of affine functions, and the computation of appropriate affine functions can be done by solving some Linear Programs. In what follows, we describe a general algorithm which does not necessarily need this property. What will be needed however is the following assumption which is satisfied again by the above particular sets.

Assumption 2. The subsets $\mathcal{L}_t$ of Assumption 1 are lattices for the pointwise partial order: for all $t\in[\![0,T]\!]$, and $\phi,\phi'\in\mathcal{L}_t$, there exists a supremum (least upper bound) of ϕ and ϕ' in $\mathcal{L}_t$, that we shall denote by $\phi\vee\phi'$ and an infimum (greatest lower bound) of ϕ and ϕ' in $\mathcal{L}_t$, that we shall denote by $\phi\wedge\phi'$.

Note that the set $\mathcal{C}_t$ of concave functions that are L_t-Lipschitz continuous on their compact support X_t is stable by the infimum operation, so that the pointwise infimum in the set of all functions coincides with the infimum in $\mathcal{C}_t$. It is not stable by the pointwise supremum, but the supremum in $\mathcal{C}_t$ exists and coincides with the concave hull of (the least concave map greater than) the pointwise supremum.

The Tropical Dynamic Programming (TDP) algorithm of Akian et al. (2020) (which generalizes Philpott et al. (2013); Baucke et al. (2018)) consists in the iterative construction of two approximations of the value function V_t, one from above and one from below. At each iteration k, the upper approximation, denoted $\overline{V}_t^k$, is obtained as the infimum (in $\mathcal{L}_t$) of a finite set $\overline{F}_t^k$ of basic functions and the lower approximation, denoted $\underline{V}_t^k$, is obtained as the supremum (in $\mathcal{L}_t$) of a finite set $\underline{F}_t^k$ of basic functions. Basic functions for the upper and lower approximations are taken respectively in subsets $\underline{\mathbf{F}}_t$ and $\overline{\mathbf{F}}_t$ of $\mathcal{L}_t$, that is we have $\overline{F}_t^k\subset\overline{\mathbf{F}}_t$ and $\underline{F}_t^k\subset\underline{\mathbf{F}}_t$. Note that the approximations $\overline{V}_t^k$ and $\underline{V}_t^k$ are parametrized by the sets $\overline{F}_t^k$ and $\underline{F}_t^k$, which means that we never store the values of these functions on a grid of $\mathbb{X}$. The sets of basic functions $\overline{F}_t^k$ and $\underline{F}_t^k$ are increasing with respect to iteration number

k, so that $\overline{V}_t^{k+1} \leq \overline{V}_t^k$ and $\underline{V}_t^{k+1} \geq \underline{V}_t^k$. These sets are computed using a sequence of state-action-noise, which is itself computed using the previous sequence of functions.

Starting with an initial state x_0 and emptysets, or appropriate singleton sets $\underline{\phi}_{t+1}^0$ and $\overline{\phi}_{t+1}^0$, the algorithm solving the Bellman equation in the hazard-decision framework consists at each step $k \geq 0$ in the following two phases:

- **Forward phase:** Compute a new (deterministic) trajectory $(x_t^k)_{t\in[\![0,T]\!]}$ starting in x_0 as follows. For $t = 0, \dots, T-1$, do:

  For each $w \in \mathbb{W}$, compute an optimal control u_t^w for $\overline{V}_{t+1}^k$ at x_t^k:
  $$u_t^w \in \arg\max_{u\in\mathbb{U}} \mathcal{Q}_t(\overline{V}_{t+1}^k)(x_t^k, u, w) \ . \tag{2}$$
  Compute the noise $w_t \in \mathbb{W}$ which maximizes the future gap
  $$w_t \in \arg\max_{w\in\mathbb{W}} \big(\overline{V}_{t+1}^k - \underline{V}_{t+1}^k\big)\big(f_t^w(x_t, u_t^w)\big).$$
  Compute the next state associated to the above noise and optimal control:
  $$x_{t+1}^k = f_t^{w_t}(x_t^k, u_t^{w_t}) \ .$$
- **Backward phase:** For $t = T, T-1, \dots, 0$, select for both upper and lower approximations, one new basic function $\overline{\phi}_t \in \overline{\mathbf{F}}_t$ (resp. $\underline{\phi}_t \in \underline{\mathbf{F}}_t$) and add it to the corresponding set: $\overline{F}_t^{k+1} := \overline{F}_t^k \cup \{\overline{\phi}_t\}$ and $\underline{F}_t^{k+1} := \underline{F}_t^k \cup \{\underline{\phi}_t\}$.

  If $t = T$, the new basic functions are chosen such that
  $$\overline{\phi}_T \geq \psi \quad \text{and} \quad \overline{\phi}_T(x_T^k) = \psi(x_T^k).$$
  and symmetrically
  $$\underline{\phi}_T \leq \psi \quad \text{and} \quad \underline{\phi}_T(x_T^k) = \psi(x_T^k).$$
  If $t < T$, the new basic functions are chosen such that
  $$\overline{\phi}_t \geq \mathfrak{B}_t\left(\overline{V}_{t+1}^{k+1}\right)$$
  $$\overline{\phi}_t(x_t^k) = \mathfrak{B}_t\left(\overline{V}_{t+1}^{k+1}\right)(x_t^k).$$
  and symmetrically
  $$\underline{\phi}_t \leq \mathfrak{B}_t\left(\underline{V}_{t+1}^{k+1}\right)$$
  $$\underline{\phi}_t(x_t^k) = \mathfrak{B}_t\left(\underline{V}_{t+1}^{k+1}\right)(x_t^k),$$
  where for all t, k, we denote $\overline{V}_t^k = \inf \overline{F}_t^k$ and $\underline{V}_t^k = \sup \underline{F}_t^k$.

For the decision-hazard framework, the only difference is in the forward phase, in which one computes an optimal control u_t independent of w:
$$u_t \in \arg\max_{u\in\mathbb{U}} \mathcal{E}_t^{(1)}(\mathcal{Q}_t(\overline{V}_{t+1}^k))(x_t^k, u) \ .$$
If $\mathcal{L}_t$ is the set of L_t-Lipschitz continuous functions on their compact support X_t, for some given norm $\|\cdot\|$ of $\mathbb{R}^n$, then a typical example of a set of basic functions $\underline{\mathbf{F}}_t$ is the set of functions $x \in X_t \mapsto a - L_t\|x - x_0\|$ with $a \in \mathbb{R}$ and $x_0 \in X_t$. Then $-\underline{\mathbf{F}}_t$ is also a good candidate for $\overline{\mathbf{F}}_t$. If $\mathcal{L}_t$ is the set of concave L_t-Lipschitz continuous functions on their compact support X_t, then one can replace $\overline{\mathbf{F}}_t$ by the set of L_t-Lipschitz continuous affine maps restricted to X_t. This is what is done in the SDDP like algorithms of Philpott et al. (2013); Baucke et al. (2018).

Theorem 3. (Akian et al. (2020)). Let V_t be the solution of the Bellman equation (1). For all $t \in [\![0, T]\!]$, the sequences $\left(\underline{V}_t^k\right)_{k\in\mathbb{N}}$ and $\left(\overline{V}_t^k\right)_{k\in\mathbb{N}}$ converge uniformly to two functions $\underline{V}_t^*$ and $\overline{V}_t^*$ of $\mathcal{L}_t$ which satisfy $\underline{V}_t^* \leq V_t \leq \overline{V}_t^*$. Moreover, we have that $\overline{V}_t^*(x_t^*) = V_t(x_t^*) = \underline{V}_t^*(x_t^*)$ for every accumulation point x_t^* of the sequence $(x_t^k)_{k\in\mathbb{N}}$. In particular $\overline{V}_0^*(x_0) = V_0(x_0) = \underline{V}_0^*(x_0)$.

The above algorithm and theorem are defined and stated in Akian et al. (2020) under some technical assumptions which ensure that Assumptions 1 and 2 hold for the set $\mathcal{L}_t$ of L_t-Lipschitz continuous functions with compact support X_t. However, the algorithm and proof only use the properties stated in these assumptions.

3. POINT BASED ALGORITHMS FOR POMDP

One way to solve a partially observable Markov decision Problem (POMDP) is to introduce, for each time t, the belief state b_t that is a probability distribution among the elements of the state space, given the information available at time t. The value function and an optimal strategy can then be obtained by solving the dynamic programming equation of a Markov decision process over the belief state space with perfect information. This gives in particular an optimal strategy which only depend on the belief state at the current time.

We recall this dynamic programming equation in the discounted infinite horizon case, for which point based algorithm were introduced (see Pineau et al. (2003); Kurniawati et al. (2008); Shani et al. (2013)). Assume that the state space is equal to $[n] := \{1, \dots, n\}$, so that the belief space is the simplex $\Delta_n = \{b \in \mathbb{R}_+^n \mid \sum_i b_i = 1\}$, which is a compact subset of $\mathbb{X} = \mathbb{R}^n$. Assume also that the observation space $\mathbb{O}$ and the control space $\mathbb{U}$ are finite sets. For all $o \in \mathbb{O}$ and $u \in \mathbb{U}$, let us denote by $\mathcal{M}^{u,o}$ the $n \times n$ matrix with entries
$$\mathcal{M}_{xx'}^{u,o} = P(\mathbf{o_{t+1}} = o, \mathbf{X_{t+1}} = x' \mid \mathbf{X_t} = x, \mathbf{U_t} = u) \ ,$$
and let us any belief state as a row $1 \times n$ vector. Then, the dynamics of the belief state is given by:
$$\mathbf{b_{t+1}} = \tau^{\mathbf{o_{t+1}}}(\mathbf{b_t}, \mathbf{u_t}) \quad \text{with} \quad \tau^o(b, u) = \frac{b\mathcal{M}^{u,o}}{b\mathcal{M}^{u,o}\mathbf{1}} \ .$$
We also have
$$P(\mathbf{o_{t+1}} = o \mid \mathbf{b_t} = b, \mathbf{U_t} = u) = p^{b,u}(o) := b\mathcal{M}^{u,o}\mathbf{1} \ .$$
Denoting $\gamma < 1$ the discount factor, the dynamic programming equation of the POMDP is the fixed point equation
$$V = \mathfrak{B}(V)$$
with $\mathfrak{B} = \mathcal{M} \circ \mathcal{E} \circ \mathcal{Q}$ for the following operators
$$\mathcal{Q}(\phi)(b, u, o) = R(b, u) + \gamma\phi\big(\tau^o(b, u)\big),$$
$$\mathcal{E}(Q)(b, u) = \sum_{o\in\mathbb{O}} p^{b,u}(o) Q(b, u, o),$$
$$\mathcal{M}(Q)(b) = \max_{u\in\mathbb{U}} Q(b, u) \ ,$$
in which $R(b, u) = b\, r(\cdot, u) = \sum_{x\in[n]} b_x r(x, u)$. The above Bellman operator has same form as the one of previous section in the decision-hazard framework, but for some special dynamics. The observation process $\mathbf{o_t}$ play the role

of the noise process $\mathbf{W_t}$. Both belong to finite sets. So we can apply the TDP algorithm as soon as we found some appropriate sets $\mathcal{L}_t$. It is well known that the value function of a POMDP is a bounded convex Lipschitz continuous function over the simplex Δ_n. The bound and the Lipschitz constant (with respect to the ℓ_1 norm on the simplex) are both equal to $L = R_{\max}/(1-\gamma)$, where $R_{\max}$ is the sup-norm of the reward function r. The point based algorithms developped in (Pineau et al. (2003); Kurniawati et al. (2008); Shani et al. (2013)) consist in approximating the value function from below by a supremum of linear maps and from above by either an infimum of functions of the form $b \mapsto a + L\|b - b_0\|_1$ or by the convex hull of such functions. Both methods can be seen as a particular case of the TDP algorithm, up to some improvements, and a generalization to the infinite horizon case. Such a generalization consists in gathering (at each iteration k) all the improvements into the same approximate value function $\overline{V}^k$ or $\underline{V}^k$, and in stopping the trajectory (x_t^k) at a time T such that $\left(\overline{V}^k - \underline{V}^k\right)(x_T^k) \leq \epsilon$.

In Smith and Simmons (2005), an analysis of the point based algorithm is done, only under the assumption that the algorithm stops. Moreover, in Fehr et al. (2018), it is proved that the Bellman operator $\mathfrak{B}$ preserves the set of Lipschitz continuous functions over the simplex, but the Lipschitz constant can become very large for some $\gamma > 1/2$.

We can show however that the set $\mathcal{L}$ of functions on the simplex which are bounded by L and can be extended in a positively homogenous and L-Lipschitz continuous map on the positive cone $\mathbb{R}^n_+$ is preserved by the Bellman operator of the POMDP. This set is compact for the uniform convergence topology, so it satisfies Assumption 1. It also satisfies Assumption 2. This allows us to construct a variant of point based algorithm for which a convergence result similar to Theorem 3 up to ϵ is possible.

REFERENCES

Akian, M., Chancelier, J.P., and Tran, B. (2020). Tropical dynamic programming for lipschitz multistage stochastic programming. ArXiv:2010.10619.

Akian, M. and Fodjo, E. (2018). From a monotone probabilistic scheme to a probabilistic max-plus algorithm for solving Hamilton-Jacobi-Bellman equations. In *Hamilton-Jacobi-Bellman equations*, volume 21 of *Radon Ser. Comput. Appl. Math.*, 1–23. De Gruyter, Berlin.

Alla, A., Falcone, M., and Saluzzi, L. (2019). An efficient DP algorithm on a tree-structure for finite horizon optimal control problems. *SIAM J. Sci. Comput.*, 41(4), A2384–A2406. doi:10.1137/18M1203900.

Baucke, R., Downward, A., and Zakeri, G. (2018). A deterministic algorithm for solving stochastic minimax dynamic programmes. *Preprint, available on Optimization Online*, 36.

Bellman, R. (1984). *Dynamic Programming*. Princeton Univ. Pr, Princeton, NJ.

Falcone, M. and Ferretti, R. (2014). *Semi-Lagrangian approximation schemes for linear and Hamilton-Jacobi equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Fehr, M., Buffet, O., Thomas, V., and Dibangoye, J. (2018). rho-pomdps have lipschitz-continuous epsilon-optimal value functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kurniawati, H., Hsu, D., and Lee, W.S. (2008). Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, volume 2008. Zurich, Switzerland.

McEneaney, W.M. (2007). A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs. *SIAM J. Control Optim.*, 46(4), 1239–1276. doi: 10.1137/040610830.

McEneaney, W.M., Kaise, H., and Han, S.H. (2011). Idempotent method for continuous-time stochastic control and complexity attenuation. In *Proceedings of the 18th IFAC World Congress, 2011*, 3216–3221. Milano, Italie.

Pereira, M.V.F. and Pinto, L.M.V.G. (1991). Multi-stage stochastic optimization applied to energy planning. *Math. Programming*, 52(2, Ser. B), 359–375. doi: 10.1007/BF01582895.

Philpott, A., de Matos, V., and Finardi, E. (2013). On Solving Multistage Stochastic Programs with Coherent Risk Measures. *Operations Research*, 61(4), 957–970. doi:10.1287/opre.2013.1175.

Pineau, J., Gordon, G., Thrun, S., et al. (2003). Point-based value iteration: An anytime algorithm for pomdps. In *IJCAI*, volume 3, 1025–1032.

Qu, Z. (2014). A max-plus based randomized algorithm for solving a class of HJB PDEs. In *53rd IEEE Conference on Decision and Control*, 1575–1580. doi: 10.1109/CDC.2014.7039624.

Shani, G., Pineau, J., and Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1), 1–51.

Smith, T. and Simmons, R. (2005). Point-based pomdp algorithms: Improved analysis and implementation. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI-05)*.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Linear Quadratic Control from an Optimization Viewpoint

Yujie Tang [*] **Yingying Li** [**] **Yang Zheng** [***] **Runyu Zhang** [*]
Na Li [*]

[*] *School of Engineering and Applied Sciences, Harvard University, Allston, MA 02134 USA (e-mail: yujietang@seas.harvard.edu, runyuzhang@fas.harvard.edu, nali@seas.harvard.edu).*
[**] *Coordinated Science Laboratory, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA (e-mail: yl101@illinois.edu)*
[***] *Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093 USA (e-mail: zhengy@eng.ucsd.edu)*

Keywords: Linear quadratic control, distributed reinforcement learning, zeroth-order optimization, connectivity, stationary points

AMS Classification: 49N10

Reinforcement learning for control systems with unknown or complex system models has attracted considerable attention recently. Particularly, there have been exciting advances on reinforcement learning for the (centralized) linear quadratic regulator problem that adopt an optimization viewpoint for algorithm design and theoretical analysis. Specifically, it has been shown that the LQR cost, when viewed as a function of the controller's feedback gain, is a gradient dominated function with a connected domain, and model-free reinforcement learning algorithms based on zeroth-order gradient estimation can achieve fast convergence to the globally optimal solution. Motivated by such success, this work attempts to investigate model-free reinforcement learning of two other linear quadratic control problems from an optimization viewpoint: i) decentralized linear quadratic control, ii) linear quadratic Gaussian control.

We first study distributed reinforcement learning of decentralized linear quadratic control. We consider a group of N agents interacting with a discrete-time linear dynamical system. Each agent only has access to partial state observations, local actions and local costs. The group of agents are connected by a communication network. The goal for each agent is to learn a local control policy by interacting with the linear system and by exchanging information with its neighbors in the network, so that the infinite-horizon averaged global cost will be minimized.

We propose a Zero-Order Distributed Policy Optimization algorithm (ZODPO) that learns local control policies in a distributed fashion. ZODPO leverages the ideas of policy gradient, zeroth-order optimization and consensus algorithms. ZODPO operates on two timescales: The fast timescale iterations proceed at the same pace with the discrete-time linear system, while each iteration on the slow timescale carries out one stochastic gradient descent update. At the beginning of each slow timescale iteration, each agent will generate a random perturbation and apply the perturbed control policy to the system. Then we let the system evolve for a sufficiently long period, during which each agent will accumulate local costs and run a consensus-based method on the fast timescale for estimating the global objective value. At the end of the slow timescale iteration, each agent constructs a zeroth-order partial gradient estimator and carries out a stochastic gradient descent update.

Further, we investigate the nonasymptotic performance of ZODPO for linear static local controllers. We show that, as long as the algorithmic parameters are properly chosen, all intermediate control policies $(\mathsf{K}(1), \ldots, \mathsf{K}(T_G))$ generated by ZODPO will stabilize the system with high probability. In addition, we derive a bound for the sample complexity of ZODPO, defined as the number of samples needed to approach a stationary point with arbitrary precision: In order to achieve

$$\frac{1}{T_G}\sum_{s=1}^{T_G} \|\nabla J(\mathsf{K}(s))\|^2 \leq \epsilon$$

for a sufficiently small tolerance $\epsilon > 0$, ZODPO requires a sample complexity of

$$\Theta\left(\frac{n_K^3}{\epsilon^4}\max\left\{n\beta_0^2, \frac{N}{1-\rho_W}\right\}\right).$$

Here n_K denotes the dimension of the controller parameter K, n denotes the dimension of the global state, β_0 is a constant determined by the system, and ρ_W captures the rate of consensus via the communication network. To the best of our knowledge, this is the first sample complexity

⋆ This work was supported by NSF CAREER grant ECCS-1553407, NSF AI Institute grant 2112085, and ONR YIP grant N00014-19-1-2217.

result for reinforcement learning of decentralized linear quadratic control.

We complement our theoretical results with numerical experiments on a multi-zone HVAC system test case.

Note that our theoretical results for ZODPO can only guarantee that a stationary point can be approached. This limitation is closely related to the fact that the optimization landscape of partially observable linear quadratic control with static output feedback lacks good structural properties that can facilitate convergence to the globally optimal point. This motivates us to analyze the optimization landscape of the linear quadratic Gaussian problem, which considers the optimal control of partially observable linear systems by dynamic controllers.

Specifically, we adopt an optimization viewpoint and reformulate the continuous-time linear quadratic Gaussian problem as an optimization problem. We parametrize a dynamic controller by its system matrices $(A_\mathsf{K}, B_\mathsf{K}, C_\mathsf{K})$. The objective function is the infinite-horizon quadratic cost, and the feasible region is the set of full-order dynamic controllers that can internally stabilize the plant.

We first characterize the connectivity of the feasible region of the LQG optimization problem. We prove that the feasible region can be disconnected, but has at most two path-connected components. Moreover, when the feasible region is disconnected, its two path-connected components are diffeomorphic under a similarity transformation $(A_\mathsf{K}, B_\mathsf{K}, C_\mathsf{K}) \mapsto (TA_\mathsf{K}T^{-1}, TB_\mathsf{K}, C_\mathsf{K}T^{-1})$ for any invertible T with $\det T < 0$, and this similarity transformation also preserves the objective value. This brings positive news to gradient-based local search algorithms for the LQG problem, since it makes no difference to search over either path-connected component even if the feasible region is disconnected. We further show that if the plant can be stabilized by a reduced-order dynamic controller, then the feasible region is always connected; this sufficient condition for connectivity becomes necessary when the plant is single-input or single-output.

We then investigate structural properties of the stationary points and the globally optimal points of the LQG cost function. It is known that the LQG cost is invariant under similarity transformations on the dynamic controller. As a consequence of this symmetry, the globally optimal solutions to the LQG problem are not unique, not isolated, and can be disconnected in the state-space domain. For a class of LQG problems, we show that the set of globally optimal solutions forms a submanifold with two path-connected components. When characterizing the set of stationary points, the notion of minimal controllers (controllable and observable controllers) plays an important role. We show it is likely that there exist many strictly suboptimal stationary points of the LQG cost function, and these stationary points are always non-minimal. We even provide an example of a saddle point that has a vanishing Hessian. In contrast, we prove that all minimal stationary points are globally optimal solutions to the LQG problem. These minimal stationary points are identical up to similarity transformations. This is expected from the classical result that the globally optimal LQG controller is unique in the frequency domain. Our analysis implies that if local search iterates converge to a critical point that corresponds to a controllable and observable controller, then the algorithm has found a globally optimal solution to the LQG problem. These results reveal rich yet complicated optimization landscape properties of the LQG problem, and shed light on the algorithm design and performance analysis of model-free policy gradient methods for solving the LQG problem.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Application of Generalized Functions in Optimal Control

Erik I. Verriest *

* *Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: erik.verriest@ece.gatech.edu).*

Abstract: A new spin is given on the classical optimal control problem with piecewise differentiable dynamics and performance index with respect to the state variables. While in each domain of differentiability, the necessary conditions for optimality are easily established, their interpretation at the boundaries between domains is not well-understood. In this paper we show that in order to make sense of the Euler-Lagrange equation at this interface one needs to transcend the classical theory of Schwartz distributions and make suitable extensions to allow for the questionable behavior of impulses multiplied by discontinuities, and the notion of partial derivatives at a discontinuity. Such a theory has been developed, in the Colombeau, Oberguggenberger and Rosinger theory of Generalized Functions in 1990, going back to ideas from Nonstandard Analysis (NSA). We develop an alternative NSA based approach applicable to impulsive dynamics and optimal control.

Keywords: Generalized Functions, Optimal Control

1. INTRODUCTION

At the Boulder AMS Conference in 1990, R. Hermann presented a talk on the Colombeau, Oberguggenberger and Rosinger theory of generalized functions and predicted that it would revolutionize mathematical physics and system theory in the 21-st century (Hermann (1994)). This has not happened, perhaps because the details involved on the extended Schwartz distribution theory are not very transparent. In this paper, the theory is revisited, and it is shown that it has indeed a distinctive merit.

Assume that the state space, for simplicity embedded in $\mathbb{R}^n$, is partitioned in two domains, Ω_+ and Ω_-, separated by the interface $\Omega_0 = \{\, x \,|\, g(x) = 0 \,\}$. A standard optimal control problem is to find the control and trajectory for a system governed by the dynamics $\dot{x} = f(x,u)$, in going from an initial state, x_0, to a desired final state, x_f, while minimizing some performance index, say $J = \int_0^T L(x,u)\,\mathrm{d}t$, with T fixed for simplicity. In the hybrid problem, L and f are smooth functions of the arguments in each domain, but are discontinuous at Ω_0. The optimal trajectory from a given state $x_0 \in \Omega_-$ to a given final state $x_f \in \Omega_+$ must cross the interface an odd number of times. Let's focus here on the case having one such crossing. Transversality requires that $[\nabla g f_-]_{t_-}[\nabla g f_+]_{t_+} > 0$, where $t_- = t_i - \epsilon$ and $t_+ = t_i + \epsilon$ for $\epsilon \xrightarrow{>} 0$. Crossing time and crossing state are a priori unknown.

Classically, the solution is found stepwise: First, establish the necessary conditions for optimality in Ω_-: The optimality condition relates u_- to x_- and λ_- implicitly. In principle one can express u explicitly so that the state equation and the Euler-Lagrange equation (EL), $\dot{\lambda}_- = -\left(\frac{\partial H_-}{\partial x}\right)^\top$, where $H_-(x,u,\lambda) = L(x,u) + \lambda^\top f$ is the Hamiltonian, leaves $2n$ coupled ODE's for x_- and λ_-. The initial condition, x_0, provides n initial conditions for these equations. Augmenting with the n unknown parameters, $\lambda_0 = \lambda_-(0)$, the solution of these $2n$ ODE's can be specified in parameterized form as $x_-(t;\lambda_0)$, with $x_-(0;\lambda_0) = x_0$, for the states, and the co-states $\lambda_-(t;\lambda_0)$. Likewise, in the domain Ω_+, the necessary conditions lead to the $2n$ coupled ODE's, parameterized by say $\lambda_+(t_f) = \lambda_f$, which are as yet unknown. This leads to the parameterized form of the n states $x_+(t) = x(t;\lambda_f)$ and the n co-states $\lambda_+(t;\lambda_f)$.

This yields two families of trajectories: One starting from x_0 and parameterized by n parameters in λ_0. The other family, $x_+(t;\lambda_f)$, ends at x_f and is likewise parameterized by the n parameters in λ_f. Continuity of the state at the interface requires that for some time τ, $x_-(\tau;\lambda_0) = x_+(\tau;\lambda_f)$, which provides n equations, but introduces another variable: τ. However, at the interface $x_i = x_-(\tau) = x_+(\tau) \in \Omega_0$ the equation $g(x_-(\tau)) = 0$ determines this τ in principle. This still leaves n unknowns in the problem. One solves for these n remaining parameters by *optimization*.

But, a close inspection of the problem reveals that the *information present in the Euler-Lagrange equation of the optimal problem has not been exhausted*. By separately solving the EL equation in both domains, the EL equation *across* the interface has not been used. Since λ_- and λ_+ may differ on $g(x) = 0$, it means that the derivative of λ must be impulsive across the interface. Thus the right-hand-side of the EL is impulsive as well. But there are several problems with this. The right-hand side of the EL equation is specified as the gradient with respect to x. If x were smooth, Schwartz's distribution theory tells us how to relate the impulse with argument x to the impulse with argument t, but here $\dot{x}$ is not continuous across the interface. Likewise the right hand side also involves a product of distributions with respect to x with functions that are necessarily discontinuous. Such objects are not allowed in

Schwartz's distribution theory. This is the essence of the *Schwartz impossibility theorem*, relating to the fact that the set of distributions does not have the structure of an algebra. See Oberguggenberger (1992). Consequently, the needed interpretation of the EL completely fails in the classical Schwartz sense at the interface.

2. KRYLOV-SPACE

The way out of this impasse is to use the framework of generalized functions with a well-defined multiplicative structure, such as outlined in Grosser et al (2001). We propose a new definition of generalized functions, one that can handle the interpretative problems of $\delta(x(t))$ and $\phi(t)\delta(t)$ respectively when x is not differentiable and ϕ is not continuous. Non-standard analysis (NSA) provides an answer, but seems more complicated than it needs to be (Goldblatt (1998). In previous work Hyun and Verriest (2016, 2017) we developed an axiomatic approach from the ground up. This was introduced for the purpose of placing *causality*, understood in the sense of a cause leading to an effect, back in hybrid system theory. While in the linear theory, the reachability problem is investigated using impulsive inputs (the cause) creating jumps in the state (the effect), in the nonlinear case hybrid systems are modeled at their jumps solely by the effects, e.g., $x(t_+) = \phi(x(t_-), p)$, where the p may be some control parameters. In the cited thesis, the objective was to put the impulses back in the continuous dynamical equation, but properly model the dynamic behavior in our sense. This theory borrowed also from the generalized function theory of Colombeau (1985), which has found widely acceptance in mathematical physics. See also Oberguggenberger (1992).
Within our *post-Schwartzian generalized function theory*, we were able to obtain a clear understanding of this singular behavior of the EL equation $\dot{\lambda} = -H_x^\top$, across the interface, which we called the *generalized Euler-Lagrange equation* (GEL). The results then lead to a precise formula for the jump $\Delta\lambda$ in the co-states across the interface $g(x) = 0$ in terms of the co-states just before and just after the Ω_0 crossing (Zhou and Verriest (2022)). The upshot of this is now that the *GEL provides n additional equations* precisely in the remaining parameters. Hence the GEL can be used instead of solving the classical parameter optimization problem. Solving nonlinear equations is computationally more straightforward. This information has not been exploited in problems involving singularities in optimal control before.

3. SCHWARTZ DISTRIBUTION THEORY AND BEYOND

The Schwartz distribution theory centers on $\mathcal{D} = C_0^\infty(\mathbb{R})$, the space of C^∞ functions with compact support. The space of distributions is then defined as the dual space $\mathcal{D}'$. A distribution $\tilde{f} \in \mathcal{D}'$ is *regular* if there exists $f \in L_{1,loc}$, the space of locally integrable functions, such that

$$\forall \phi \in \mathcal{D}, \quad \tilde{f}(\phi) = \int_{\mathbb{R}} f(x)\phi(x)\,\mathrm{d}x.$$

Let $\mathcal{D}'_{reg}$ denote the space of *regular distributions*. Then $\mathcal{D}'_{reg} \subset \mathcal{D}'$ and $\forall f, g \in L_{1,loc}$ it holds that $\tilde{f} = \tilde{g} \Leftrightarrow f = g$ a.e.. In contrast, the *evaluation functional*, $\sigma_t : C_0^\infty \to \mathbb{R} : \phi \mapsto \phi(t)$, is usually denoted as $\int_{\mathbb{R}} \delta_t(x)\phi(x)\,\mathrm{d}x$, however the δ_t known as Dirac-delta's are not functions.
For $D \in \mathcal{D}'$, the *distributional derivative* is defined by $D' : \mathcal{D}' \to \mathcal{D}' : D \to (\phi \mapsto -D(\phi'))$. For $\tilde{f} \in C^1 \cap L_{1,loc}$, we get $\widetilde{(f')} = (\tilde{f})'$. For smooth functions $f \in C^\infty$ and $D \in \mathcal{D}'$, we define $fD : C_0^\infty \to \mathbb{R} : \phi \mapsto D(f\phi)$, and note that $fD \in \mathcal{D}'$. Furthermore, $(fD)' = f'D + fD'$, and $\forall g \in L_{1,loc}$, $f\tilde{g} = \widetilde{(fg)}$.
The *support of a distribution* is the complement of the largest open set on which the distribution vanishes. Let $\mathcal{D}'_M = \{D \in \mathcal{D}' \operatorname{supp} D \subseteq M\}$ for any measurable subset of $\mathbb{R}$. Let D_M denote the *restriction* of the distribution D. Desirable properties for a restriction are:
i) $\forall D \in \mathcal{D}'$ and measurable M, $D_M \subseteq \mathcal{D}'_{\mathrm{cl}\,M}$, and D_M linear and idempotent.
ii) $\forall f \in L_{1,loc}$ and measurable M, if $f_M = \chi_M f$, with χ_M the indicator function of M, then $\widetilde{f_M} = (\tilde{f})_M$.
iii) $\forall \phi \in C_0^\infty$, $\forall D \in \mathcal{D}'$ and measurable M it holds that $\operatorname{supp}\phi \subseteq M \Rightarrow D_M(\phi) = D(\phi)$ and $\operatorname{supp}\phi \cap M = \emptyset \Rightarrow D_M(\phi) = 0$.
iv) M_i pairwise disjoint with $M = \cup M_1$, it holds that $D_M = \sum D_{M_i}$ and $(D_{M_i})_{M_j} = 0$ if $i \neq j$.
If $\operatorname{supp} D = \{t\}$, then there exists $N \in \mathbb{N}$, and $\alpha_i \in \mathbb{R}$; $i = 0, \ldots, n$ such that $D = \sum_{i=0}^N \alpha_i \delta_t^{(i)}$ and $\sum_{i=0}^N \alpha_i \delta_t^{(i)} = 0 \Leftrightarrow \alpha_i = 0$, for $i = 0, \ldots, N$.
As the conditions (i) to (iv) cannot be satisfied together, Trenn (2009) defined an appropriate subspace of *piecewise regular distributions* that allows (i) to (iv) to be satisfied.

$$\mathcal{D}'_{pw\,reg} = \{\tilde{f} + \sum_{t\in T} D_t \,|\, f \in L_{1,loc}, T \subset \mathbb{R}, \text{locally finite}\},$$

where $D_t \in \mathcal{D}'_{\{t\}}$. It is shown in (Trenn (2009)) that $\mathcal{D}'_{pw\,reg} \subset \mathcal{D}'$, the representation is unique, (i) to (iv) are satisfied, and for measurable $M \subset \mathbb{R}$, $D = \tilde{f} + \sum_T D_t$ implies $D_M = \widetilde{f_M} + \sum_T \chi_M(t)\,\mathrm{d}t$.
For $D \in \mathcal{D}'_{pw\,reg}$, define $D[t] = D_{\{t\}}$ as D_t if $t \in T$ and zero else. Then $D[\cdot] = \sum_T D_t$ is the *impulsive part* of the distribution D.
Multiplication with piecewise smooth functions is well-defined (Trenn (2009)).
The class of piecewise smooth distributions is then defined and with the differentiation and multiplication defines an associative differential algebra. In Trenn (2009) it was argued that $\widetilde{\chi_{[t,\infty)}}\,\delta_t = \delta_t$ for all t corresponds with a causality condition, rendering the Cauchy initial value problem of an algebraic differential equation unique. While successful in switched behaviors with impulses (Trenn and Willems (2012)), the application of this multiplication to the Euler-Lagrange equation in multi-mode optimal control is inconsistent with the solution method that computes the set of optimal solutions in each domain, parameterized by the interface condition, and then determines the optimal parameters. This stems from the fact that the Euler-Lagrange and state equations constitutes a two-point boundary value problem and not an initial value problem. Hence *causality is not a concern*, and the solutions in the above extension of $\mathcal{D}'_{pw\,reg}$ will not solve the problem. The way out is to enlarge the set of distributions in the nonstandard sense. Indeed, Todorov (1990) proved the existence of a nonstandard function ${}^*\delta \in {}^*C_0^\infty$, such

that for all $\phi \in C^0$

$$\int_{^*\mathbb{R}} {}^*\delta(x)\, {}^*\phi(x)\, \mathrm{d}x = \phi(0).$$

This means that pointwise evaluation of a generalized distribution is possible. The next section uses some notions of non-standard analysis, see Goldblatt (1998). Essentially, just as Cantor's description of the reals is done by considering equivalence classes of Cauchy sequences of rationals, the hyperreals or non-standard reals are definable as equivalence classes (under a different equivalence relation) of sequences of reals. A sequence in $\mathbb{R}^{\mathbb{N}}$ will be denoted by $\{r_n\}_n$ or $\{r_n\}$.

4. KRYLOV SPACE

We review the formal construction in Hyun and Verriest (2017).

4.1 Krylov Hyperreals

The set of hyperreals, $^*\mathbb{R}$, defined in NSA is too large to be easily manageable. The redundancy of infinitesimals prevents the space from having a constructive property such as having a countable basis generating the hyperreals as a vector space over a field, $\mathbb{R}$. Therefore, we construct a countably infinite basis, which generates a reduced extension of the reals, denoted as $\mathbb{K}$.

Definition 1: Let $\alpha > 1$. A sequence $\langle r_n \rangle$ is called a $\mathbb{K}_i$-sequence if there exists $s \in \mathbb{R}$ such that $r_n = (\frac{1}{\alpha^{in}})s$ for all $n \in \mathbb{N}$.
This geometric form of the sequence is reminiscent of linear algebra where it would be called a Krylov sequence. We'll denote it by $\langle \frac{1}{\alpha^i}; s \rangle$. A set $\{x \in \mathbb{R} \,|\, x \text{ is a } \mathbb{K}_i\text{−sequence}\}$ is called $\mathbb{K}_i$ space. Thus $x \in \mathbb{K}_i$ means there exists $s \in \mathbb{R}$ such that $x = \langle \frac{1}{\alpha^i}; s \rangle$.
$\mathbb{K}_i$ is a one-dimensional vector space over $\mathbb{R}$ with property

$$\mathbb{K}_0 = \mathbb{R}, \quad \mathbb{K}_i \subset \mathrm{hal}(0),\ \forall i \in \mathbb{N} \setminus \{0\},$$

where $\mathrm{hal}(0) = \{\, b \in {}^*\mathbb{R} | b \text{ is infinitesimal} \,\}$ is the *halo* of zero, as defined in NSA. A convenient basis for $\mathbb{K}_i$ is the element $e_i = \left\langle (\frac{1}{\alpha})^i; 1 \right\rangle$. With the product defined above, we have $e_i \cdot e_j = e_{i+j}$ for all $i, j \in \mathbb{Z}$. With addition, the set $\mathbb{K}^N = \{x \in {}^*\mathbb{R} \,|\, x = \sum_{i=0}^{N} s_i e_i,\ \{s_n\}_{i=0}^{N} \subset \mathbb{R}\}$ is the direct sum vector space $\oplus_{i=0}^{N} \mathbb{K}_i$ with basis $\{e_i\}_{i=0}^{N}$.
The set $\mathbb{K} = \{x \in {}^*\mathbb{R} \,|\, x = \sum_{i=0}^{\infty} s_i e_i,\ \{s_n\}_{i=0}^{\infty} \in \ell_1\}$ is called the Krylov hyperreal space. In fact $\mathbb{K}$ is also a commutative ring, and the representation of $x^* \in \mathbb{K}$ with coordinates s_i is unique up to equivalence in $^*\mathbb{R}$. There are elements in the hyperreal space that are not elements of $\mathbb{K}$. However, the Krylov space is large enough to contain the reals and the infinitesimals. Given $t^* \in \mathbb{K}$ with representation $\{s_i\}_i \in \ell_1$, we shall refer to s_0 as the *sensible* time of t^*, and s_k, $k > 0$ as the *insensible* time of t^* with $1/\alpha^k$ as convergence rate.

4.2 Krylov Functions

Definition 2: A hyperreal function is a mapping $F : {}^*\mathbb{R} \to {}^*\mathbb{R}$ such that there exists a sequence $\{f_n\}_n \subset \mathbb{R}^{\mathbb{N}}$ an $t^* \in {}^*\mathbb{R}$ and $\sigma_{t^*} F \stackrel{\text{def}}{=} \langle f_n(t_n) \rangle$.
The restriction of F to the Krylov space $\mathbb{K}$ is called a *Krylov function*. We start with sequences of smooth real valued functions of the form $\langle h, S_\alpha h, S_\alpha^2 h, \ldots \rangle$ denoted as $\langle S_\alpha : h \rangle$, where $h : \mathbb{R} \to \mathbb{R}$ and S_α is the scaling operator, $\sigma_t(S_\alpha x) = \sigma_{\alpha t} x$, and take the closure under componentwise addition and multiplication

$$\langle S_\alpha; f \rangle + \langle S_\beta; g \rangle = \langle f + g, S_\alpha f + S_\beta g, S_\alpha^2 f + S_\beta^2 g, \ldots \rangle$$
$$\langle S_\alpha; f \rangle \cdot \langle S_\beta, g \rangle = \langle fg, S_\alpha f \cdot S_\beta g, S_\alpha^2 f \cdot S_\beta^2 g, \ldots \rangle.$$

Let σ_t be the evaluation functional on the class of regular functions $\forall t \in \mathbb{R} : \sigma_t h = h(t)$. The evaluation of a hyperfunction $\langle h_n \rangle$ at $t^* = \langle \gamma; t \rangle \in {}^*\mathbb{R}$ is the hyperreal $\langle h_1(t), h_2(\gamma t), h_3(\gamma^2 t), \ldots \rangle$ The Krylov space is the space of (equivalence classes) of such functions. A regular function, h is embedded in the space generated by Krylov sequences as the class of $\langle 1; h \rangle$. The idea is that a function of the form $f^* = \langle 1; f \rangle + \langle S_\alpha; g \rangle$ is a hyperfunction converging to f in a precise way determined by the rate α and shape g. The above defined hyperreal valued function is said to be in the halo of f, while f is the shadow of f^*. More generally, we have $\sigma_{\langle \beta; t \rangle} \langle f_0, f_1, f_2, \ldots \rangle = \langle f_0(t), f_1(\beta t), \ldots \rangle$, and $\sigma_{\langle 1; t_0 \rangle + \langle \beta; s_0 \rangle} \langle 1; f \rangle = \langle f(t_0 + s_0), f(t_0 + \beta s_0), f(t_0 + \beta s_0^2), \ldots \rangle$. The right limit of a regular function at t_0 is the shadow, $f(t_0+)$, if $s_0 > 0$ and $0 < \beta < 1$.

Consider now $u \in C^1(-1, 1))$, with $\lim_{t \to -1} u(t) = 0$, $\lim_{t \to 1} u(t) = 1$, and extend it to a $C^0(\mathbb{R})$ function, h, with $h(t) = 0$ for $t < -1$ and $h(t) = 1$ for $t > 1$. Then h is a model for the Heaviside function, which is defined as the generalized function $H_h = \langle S_\alpha; h \rangle$. If $t \in \mathbb{R}$, then $H_h(t)$ is the classical Heaviside function for all h. Evaluation at the infinitesimal time $\langle \frac{1}{\alpha}; \overline{t} \rangle$ gives $\sigma_{\langle \frac{1}{\alpha}; \overline{t} \rangle} \langle S_\alpha; h \rangle = \langle h(\overline{t}), \ldots, h(\overline{t}), \ldots \rangle = \langle 1; h(\overline{t}) \rangle$. On a microscopic scale, there are infinitely many Heaviside functions. Consequently, H_h^2 and H_h differ in insensible time, as they respectively give $h^2(\overline{t})$ and $h(\overline{t})$.
Similarly we can define H_h^+ by letting $h(t) = 0$ for $t \in (-1, 0]$ and $h(t) = 1$ for $t \in [0, 1)$. The derivative of H_h^+ is represented by the sequence $DH_h^+ = \langle \alpha; 1 \rangle \langle S_\alpha; Dh \rangle$. With $Dh(t) = h(t) - h(t-1)$, we see that in turn this can be made continuous by a regularization $\delta_{reg} = \langle S_\alpha; Dh \rangle_{reg} = \langle \alpha, 1 \rangle \langle S_\alpha; h_1 \rangle$ with $h_1 = S_\alpha(I - T_{-\alpha})h$. In turn, this gives for $D\delta_{reg}$ the k-th term in the sequence $DS_{\alpha^k}(I - T_\alpha)h = \alpha^k S_{\alpha_k}(I - T_{-\alpha})Dh$. Taking again the regularization $Dh = S_\alpha(I - t_{-\alpha})h$, we find $(D\delta_{reg})_{reg} = \langle \alpha^2; 1 \rangle \langle S_\alpha; S_\alpha(I - T_{-\alpha^2})(I - T_{-\alpha})h \rangle$. It is easily shown by induction that the regularized ℓ-th derivative of the delta corresponds to the generalized function $\langle \alpha^\ell; 1 \rangle \langle S_\alpha; S_{\alpha^\ell} \prod_{i=1}^{\ell} (I - T_{-\alpha^i})h \rangle$.

Alternatively, let $u \in C^\infty((0, 1))$ with again $u(1-) = 1$ and $u(0+) = 0$, and extend it to a Heaviside function. All derivatives are defined and regularization is no longer necessary. In this case the ℓ-th derivative of H_h is represented by $\langle \alpha^\ell; 1 \rangle \langle S_{\alpha^\ell}; D^\ell h \rangle$. Its evaluation at $^*t = \langle \frac{1}{\alpha^k}; \overline{t} \rangle$ is the smooth function $h(\overline{t})$ at level $k = \ell$, and 0 for the other cases.

5. MAIN RESULTS

Let $fg \in C^0$ and $F(t) = f(t) + g(t)H(t)$. The product $F(t)\delta$ is not defined as a Schwartzian distribution, but does in the Krylov sense.

Definition 3 The *impulsivity* of a generalized function ${}^*f$ at a point $t_0 \in \mathbb{R}$ is given by the integral $\int_{t_0-\langle\frac{1}{\alpha};1\rangle}^{t_0+\langle\frac{1}{\alpha};1\rangle} {}^*f(t)\,\mathrm{d}t = \langle \ldots, f_k(\frac{t-t_0}{\alpha^k})\,\mathrm{d}t, \ldots\rangle$.
For δ represented by $\langle\alpha;1\rangle\langle S_\alpha h'\rangle$, we find

$$\int_{\langle\frac{1}{\alpha};1\rangle}^{\langle\frac{1}{\alpha};1\rangle} {}^*\delta(t)\,\mathrm{d}t = \left\langle \ldots, \int_{-\frac{1}{\alpha^k}}^{\frac{1}{\alpha^k}} \alpha^k h'(\alpha^k t)\mathrm{d}t, \ldots \right\rangle = h(1) - h(-1) = 1.$$

It is well known that f is continuous at 0, then $f(t)\delta(t) = f(0)\delta(t)$ for any representation of the Heaviside function.

Theorem 1 The impulsivity of $f(t)\delta_h(t)$ with $f(t) \in C^0_{pw}$ is given by $\langle f\rangle_0 \stackrel{\text{def}}{=} \frac{1}{2}(f(0_-) + f(0_+))$.
Proof: It holds that there exists g and $k \in C^0$ and some $0 < \epsilon \in \mathbb{R}$ such that $f(t) = g(t) + k(t)H_h(t)$. Thus the impulsivity of $f(t)\delta_h(t)$ follows from

$$\begin{aligned}
&\int_{\langle\frac{1}{\alpha};1\rangle}^{\langle\frac{1}{\alpha};1\rangle} (g(t) + k(t)H_h(t))\delta_h(t)\,\mathrm{d}t \\
&= \left\langle \ldots, \int_{-\frac{1}{\alpha^k}}^{\frac{1}{\alpha^k}} \left(g(t) + k(t)h(\alpha^k t)\right)\alpha^k h'(\alpha^k t)\mathrm{d}t, \ldots \right\rangle \\
&= g(0) + \frac{1}{2}k(0)(h^2(1) - h^2(-1)) = \langle f\rangle_0.
\end{aligned}$$

Corollary On the sensible time scale $f(t)\delta(t)$ with $f \in C^0_{pw}$ is equivalent to $\langle f\rangle_0\delta(t)$, and using the shift property $f(t)\delta(t-t_0) = \langle f\rangle_{t_0}\delta(t-t_0)$.

If $x \in C^1$ with $x(t_0) = 0$ and $\dot{x}(t_0) \neq 0$, then it is known that $\delta(x(t) = \frac{\delta(t-t_0)}{|\dot{x}(t_0)|}$. The following theorem generalizes.

Theorem 2 Let $x \in C^1_{pw}(\mathcal{O}(t_0))$ where $\mathcal{O}(t_0)$ is a neighborhood of t_0, where $x(t_0) = 0$. Let x be transversal, i.e., $\dot{x}(0-)\dot{x}(0+) > 0$, then $\delta(x(t)) \approx \frac{\delta(t-t_0)}{\langle|\dot{x}|\rangle_{t_0}}$.
Here, $\langle x\rangle_{t+0}$ is the average $\frac{1}{2}(x(t_0-) + x(t_0+))$.
Proof: Use the fact that $x(t) = (t-t_0)(x_-(t)H(-t) + x_+(t)H(t))$ in a sufficiently small neighborhood of t_0 where $x_\pm \in C^1$ have the same sign. □

The solution of the ODE $\dot{x}(t) = Ax(t)\delta(t)$ is constant for $t > 0$ and $t < 0$. Its behavior at $t = 0$ is retrieved from the sequence of equations $\dot{x}_k(t) = A\alpha^k h'(\alpha^k t x_k(t))$, and leads to $x(0+) = \mathrm{e}^A x(0-)$, which is consistent with Nedeljkov and Oberguggenberger (2012), but differs from Trenn (2009).

Theorem 3 The standard optimal control problem sketched in the Introduction, with discontinuous cost rate and dynamics w.r.t. x has a jump in the costates when the interface is crossed given by

$$\Delta\lambda = -\frac{(\Delta L + \langle\lambda\rangle\Delta f)m}{\langle|m^\top f|\rangle}, \quad m = \left(\frac{\partial g}{\partial x}\right)^\top.$$

Proof: The Euler-Lagrange equation for the optimal control problem is given by

$$\dot{\lambda} = -\frac{\partial L(x,u)}{\partial x}^\top - \lambda^\top f(x,u).$$

Since L and f are discontinuous when $g(x) = 0$, the right hand side of the EL is singular (as function of x). Hence also the left hand side must have a singularity. However, the latter is with respect to the time t. It is therefore necessary to express $\delta(g(x))$ in terms of a delta with respect to time. But precisely because of this dependency, x, and thus also $g(x)$ have a discontinuous derivative when $g(x) = 0$. Thus Theorem 2 applies. Likewise, the singularity in $\frac{\partial f}{\partial x}$ is multiplied by $\lambda^\top$ which has a jump when $g(x) = 0$. Here Theorem 1 applies. Combining and integrating over time in an infinitesimal interval across the discontinuity yields the result. □
The traditional "two-domain" method of solving the problem, keeping the cross-over point and time as parameters to be optimized, is now replaced by a non-linear equation.

ACKNOWLEDGEMENTS

The author thanks the support from the Julius-Maximilans-Universität Würzburg through the Giovanni Prodi Chair.

REFERENCES

J.F. Colombeau. Elementary introduction to new generalized functions. Nort-Holland Elsevier 1985.
R. Goldblatt. Lectures on the Hyperreals: An Introduction to Nonstandard Analysis. Graduate Texts in Mathematics, Springer, New York, 1998
M. Grosser, M. Kunzinger, M. Oberguggenberger and R. Steinbauwer.
R. Herman. C-O-R generalized functions, current algebras, and control. Math Sci Press, 1994.
N.-sP.Hyun and E.I. Verriest. Causal impact modeling of state dependent impulsive affine systems using nonstandard analy6sis. Proceedings of the 55-th IEEE Conference on Decision and Control, 2016, 3024–3029.
N.-s.P. Hyun, and E.I. Verriest. A causal interpretation of nonlinear impulsive systems based on non-standard analysis. Nonlinear Analysis: Hybrid Systems 25, 138-–154.
M. Nedeljkov and M. Oberguggenberger. Ordinary Differential Equations with Delta Function Terms. Publications de 'l Institut Mathématique. Novelle série 91 (105) 125–135.
M. Oberguggenberger Multiplication of distributions and applications to partial differential equations. Pitman Research Notes in Mathematics, Vol 259. Longman 1992.
T. Todorov. A Nonstandard Delta Function Proceedings of the American Mathematical Society Volume 110, Number 4, December 1990, pp. 1143–1144.
S. Trenn. Regularity of distributional differential algebraic equations. Math. Control Signals Syst. (2009) 21:229–264.
S. Trenn and J.C. Willems. Switched behaviors with impulses - a unifying framework. Proc. 51st IEEE Conference on Decision and Control Maui, USA, December 2012, pp. 3203-3208.
M. Zhou and E.I. Verriest. Generalized Euler-Lagrange Equation: A Challenge to Schwartz's Distribution Theory. Proc. American Control Conference, Atlanta, GA June 2022.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Measures of Modal Controllability for Network Dynamical Systems

Anand Gokhale * **Manikya Valli Srighakollapu** *
Ramkrishna Pasumarthy *

* *Electrical Engineering Department, Indian Institute of Technology Madras (e-mail: ee17b158@smail.iitm.ac.in, ee16d032@smail.iitm.ac.in,ramkrishna@ee.iitm.ac.in).*

Abstract:
The quantification of controllability has gained renewed interest in the context of large, complex network dynamical systems. In some application areas such as computational neuroscience, there is a large interest in modal controllability, which describes the ability of an input to control the modes of a system. In case of a linear system, the modes of the system are given by the left eigenvectors associated with the system matrix. In this work, we identify mode specific and gross metrics for modal controllability for discrete linear time invariant systems. Our metrics are based on energy requirements to move along a given mode and find applications in problems involving selection of driver nodes for minimizing control effort along particular modes of the network. We conclude by studying the properties of the metrics.

Keywords: modal controllability, network control, large scale systems, control energy

1. INTRODUCTION

In recent years, with the development of parallel and decentralized algorithms, there has been a renewed interest in the study of network systems. Such systems appear frequently in the analysis of power grids, infrastructure networks, brain networks, and even social networks. In systems of this scale, simply answering the question of whether a system is controllable or not is not sufficient, as the energy requirements to control such a system can in some cases be impractical, especially when the size of the network is large. To analyze systems from an energy perspective, several measures have been proposed based on spectral properties of controllability gramian, such as minimum eigenvalue(Yan et al. (2012); Pasqualetti et al. (2014)), trace of inverse (Summers et al. (2016)), trace (Summers and Lygeros (2014)), and determinant(Cortesi et al. (2014)) of controllability gramian.

Modal analysis continues to be a strong tool to analyze linear systems. Classically, the study of controllability from a modal perspective involves the identification of controllable and uncontrollable modes. There are two main motivations to consider a modal analysis for large network systems. Firstly, in large networks, the system matrices may not be well-conditioned; therefore, it may be difficult to analyze the system as a whole. Such systems can be divided into modes, and each mode can be analyzed separately. Since each mode has unique conditioning, it is possible to perform a modal analysis. Secondly, in large systems, it is possible that one may be interested in particular modes of the system(for example, the unstable modes or the mode associated with the dominant eigenvalue) and not actually interested in the behavior of the entire system. The overwhelming majority of literature on this topic emphasizes the controllability/lack thereof for the system. Notably, the tests for controllability (Hespanha (2018)), such as Popov-Belevitch-Hautus (PBH) test and Eigenvector test, offer a qualitative measure of controllability. Here, we focus on a quantitative measure for modal controllability, which is important in practical situations due to the limitations of physical components in large scale networks.

To our knowledge, there are two metrics proposed in the literature to measure modal controllability. In Hamdan and Nayfeh (1989), it is proposed that the the cosine of the angle between a mode and the input vector be used as a metric. The argument to use this metric is based on the fact that an increase in the mentioned angle from 0 to 90 degrees leads to a decrease in controllability, with the system losing controllability at an angle of 90 degrees. The authors claims that the use of cosine of the angle is seen as an extension to the PBH test. While this metric accounts for the variance in modal controllability due to the angle between the mode and the input vector, it does not account for the eigenvalue associated with the mode. Due to this, this metric is unable to deal with the analysis of a set of modes. A more recent metric (Pasqualetti et al. (2014)) is based on the ability to maximize the reachability of difficult to reach states. This metric is based on a heuristic and has gained some traction in neuroscience based applications(Gu et al. (2015)).

In this abstract, we define a new metric for modal controllability based on optimal energy requirements to control a system along with a particular mode. We analyze this metric and identify the driver nodes in a network that minimize the energy requirements to move along a mode. We also identify the modes that are the easiest to control

from a given set of driver nodes. Unlike previous metrics, our metric also has a normalization factor that allows the comparison of energy requirements across modes.

The paper is organized as follows. We discuss some preliminary theory in Section 2. We propose our new metric in Section 3. In Section 4, we analyze the metric and study how it varies with varying system parameters. Finally, we conclude in Section 5.

Notation: We denote the set of real numbers by $\mathbb{R}$. $\mathbb{R}^{n\times m}$ denotes the set of matrices of dimension $n \times m$. For any matrix $A \in \mathbb{R}^{n\times m}$, $A' \in \mathbb{R}^{m\times n}$ denotes the transpose of matrix A. We also denote the j^{th} column of A as $\mathbf{a}_j$. The symbol $\mathbf{1}_n$ denotes a vector in $\mathbb{R}^n$, with all entries as 1. We define the norm of a vector $\mathbf{v} \in \mathbb{R}^n$ as $\|\mathbf{v}\| = \sqrt{\mathbf{v}'\mathbf{v}}$.

2. PRELIMINARIES

2.1 Linear Systems Theory

We represent a discrete linear time invariant system (D-LTI system) by

$$\mathbf{x}^+ = A\mathbf{x} + B\mathbf{u}. \tag{1}$$

where, $\mathbf{x} \in \mathbb{R}^n$ represents the state of a system, $A \in \mathbb{R}^{n\times n}$ is the system matrix, and $B \in \mathbb{R}^{n\times m}$ is termed as an input matrix. $\mathbf{u} \in \mathbb{R}^m$ is the control input.

In the special case of networked control systems, the input matrix may be denoted by $B_{\mathcal{K}}$, whose columns correspond to the canonical vectors associated with the input/driver nodes in the network. Further, in network systems, the matrix A represents the adjacency matrix of the network.

If the eigenvalues and both left and right eigenvectors of A can be computed, they contain a lot of information that describes the system in certain circumstances. For example, we may be interested in the dominant eigenvalue, and its associated eigenvector. In a case where the system matrix A is doubly stochastic, the dominant mode is along the vector $\mathbf{1}_n$. Here, controlling the system along this mode could be used to influence the average state of all nodes. As discussed in Bullo (2022), such systems appear in applications such as cyclic pursuit in robotic networks, wireless sensing networks and animal flocking behaviour analysis. For the purpose of modal analysis, the eigenvalues for A must not only be distinct, but A should also be well conditioned. For matrices which have a mixture of both well conditioned and ill conditioned eigenvalues, we limit our methods to the well conditioned eigenvalues.

Assuming A is sufficiently well conditioned, and has distinct eigenvalues, we consider the left eigenvectors of A as the modes of the system, similar to Hamdan and Nayfeh (1989). To guarantee uniformity across modes, we scale the left eigenvectors to ensure their norm is unity.

2.2 Controllability Gramian

The PBH test(Hespanha (2018)) provides a qualitative result, i.e. it does not quantify the difficulty in controlling a system. In large scale systems, a binary answer, of whether or not the system is controllable, may not be sufficient as there may be practical constraints on the input energy. In literature, several metrics have been proposed based on controllability gramian, to quantify the control energy (Müller and Weber (1972); Cortesi et al. (2014)). For the D-TI system described in Equation (1),the controllability gramian is given by

$$W_c(T) = \sum_{\tau=0}^{T} A^\tau BB'(A')^\tau.$$

We note that the minimum energy required to drive a linear system from the origin to a point x_f, in time horizon T is given by

$$E = \mathbf{x}_f' W_c^{-1}(T)\mathbf{x}_f,$$

2.3 Modularity

In our work, we are concerned with the selection of the set of best driver nodes to control a mode, and the set of modes that are easiest to control from a set of inputs. We formulate both problems as set function optimization problems. In this regard, we define modular set function as follows:

Definition 2.1 (Lovász (1983))**.** *Let V be a set. A set function $f : 2^V \to \mathbb{R}$ is said to be modular if and only if for any subset $S \subseteq V$, it can be expressed as*

$$f(S) = a(\phi) + \sum_{i\in S} a(i) \tag{2}$$

for some weight function $a : V \to \mathbb{R}$ and ϕ denoting the null set.

Optimization problems involving the selection of a subset $S \subseteq V$ in order to maximize or minimize a modular function can be solved via a sorting algorithm.

2.4 An extension to Cauchy Schwarz inequality

To motivate our arguments, we use an extension to Cauchy Schwarz inequality.

Lemma 2.1. *Consider a symmetric, positive definite matrix $Q \in \mathbb{R}^{n\times n}$. For any $\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| = 1$, we have*

$$\mathbf{x}'Q\mathbf{x}\mathbf{x}'Q^{-1}\mathbf{x} \geq 1.$$

Proof. Since Q is positive definite and symmetric, it can be diagonalized by an orthogonal matrix (Horn and Johnson (2012)), i.e,

$$Q = P'DP, \quad \text{and}$$
$$Q^{-1} = P'D^{-1}P.$$

where P is orthogonal. Orthogonal matrices preserve norms, so $\|P\mathbf{x}\| = 1$. Let $\mathbf{y} = P\mathbf{x}$. The LHS of our inequality is now reduced to

$$\mathbf{y}'D\mathbf{y}\mathbf{y}'D^{-1}\mathbf{y} = \sum_{i=0}^{n} d_i\mathbf{y}_i^2 \sum_{i=0}^{n} \frac{\mathbf{y}_i^2}{d_i},$$

where d_i is the i^{th} diagonal element of D.
Since Q is positive definite, $d_i > 0$, and since $\|\mathbf{y}\|_2 = 1$, $\sum_{i=1}^{n} \mathbf{y}_i^2 = 1$. Applying Cauchy-Schwarz,

$$\sum_{i=0}^{n} d_i\mathbf{y}_i^2 \sum_{i=0}^{n} \frac{\mathbf{y}_i^2}{d_i} \geq \sum_{i=1}^{n} \frac{\sqrt{d_i}}{\sqrt{d_i}}\mathbf{y}_i^2 = 1.$$

□

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

3. PROPOSED METRIC FOR MODAL CONTROLLABILITY

In this section, we propose modal controllability metrics for D-LTI systems.

Theorem 3.1. *Let us consider a discrete time LTI system* (1) *and assume* (A, B) *is controllable. Let* $\mathbf{v}_i$ *be a left eigenvector of* A*, associated with the eigenvalue* λ_i*. Let the optimal energy required to traverse one unit in the direction of* $\mathbf{v}_i$ *be* E_i*, for a time horizon* t*, starting from the origin. Then,*

$$E_i \geq \left(\mathbf{v}_i' BB' \mathbf{v}_i \frac{1-\lambda_i^{2(t+1)}}{1-\lambda_i^2} \right)^{-1}.$$

Proof. The optimal control energy for a system to reach $\mathbf{v}_i$ starting from the origin is given by

$$E_i = \mathbf{v}_i' W_c^{-1}(t) \mathbf{v}_i,$$

where W_c is the controllability gramian. Since (A, B) is controllable, the controllability gramian is symmetric and positive definite. By Lemma 2.1, we have.

$$E_i \mathbf{v}_i' W_c(t) \mathbf{v}_i \geq 1. \tag{3}$$

The controllability gramian, for a time horizon t is given by ,

$$W_c(t) = \sum_{k=0}^{t} A^k BB' (A')^k.$$

Therefore,

$$\begin{aligned} \mathbf{v}_i' W_c(t) \mathbf{v}_i &= \mathbf{v}_i' \left(\sum_{k=0}^{t} A^k BB' (A')^k \right) \mathbf{v}_i \\ &= \sum_{k=0}^{t} \mathbf{v}_i' A^k BB' (A')^k \mathbf{v}_i \\ &= \mathbf{v}_i' BB' \mathbf{v}_i \sum_{k=0}^{t} \lambda_i^{2k}. \end{aligned}$$

The theorem follows from Equation (3) and the above equation. □

Based on this theorem, we define a metric for discrete time systems. We consider the following metric M_i^d to control the i^{th} mode from the given input matrix B, for a time horizon t,

$$M_i^d(t) = \left(\mathbf{v}_i' BB' \mathbf{v}_i \frac{1-\lambda_i^{2(t+1)}}{1-\lambda_i^2} \right)^{-1}.$$

We extend this metric to quantify the difficulty of controlling multiple modes as follows:

Corollary 3.1. *For the setup described in Theorem 3.1, we have*

$$\sum_{i \in \mathcal{T}} E_i \geq \sum_{i \in \mathcal{T}} \left(\mathbf{v}_i' BB' \mathbf{v}_i \frac{1-\lambda^{2(t+1)}}{1-\lambda_i^2} \right)^{-1}.$$

The proof for this statement follows from the theorem, by summing across modes. Subsequently, we define the following metric across modes. For a given set of modes, $\mathcal{T}$ and a time horizon t, we have,

$$M^d(\mathcal{T}, t) = \left(\sum_{i \in \mathcal{T}} \frac{1}{M_i^d(t)} \right)^{-1}.$$

In the limiting case, as $t \to \infty$, when A is stable, we have,

$$\sum_{i \in \mathcal{T}} E_i \geq \sum_{i \in \mathcal{T}} \left(\mathbf{v}_i' BB' \mathbf{v}_i \frac{1}{1-\lambda_i^2} \right)^{-1},$$

where the inequality follows as finite horizon energy control is always higher than the infinite horizon case.

4. DISCUSSION

4.1 Factors affecting the metric

A lower bound for the energy required to control a mode is inversely related to $\mathbf{v}_i BB' \mathbf{v}_i$. Since the eigenvector $\mathbf{v}_i$ is normalized, this term only depends on the magnitude of the columns in the input matrix, and the angle between the column of the input matrix and the eigenvector. This is in line with the PBH test, as when the two vectors are orthogonal, the system is uncontrollable.

Further, a key feature of our metric is the ability to compare the ease of controlling a system across modes. This allows us to study optimization problems where we consider the optimization of the ease of control across a range of modes. Specifically, in the discrete case, we note that modes with eigenvalues closer to unity are easier to control. This is in part due to the natural response of the system. This result is consistent with the findings in Lindmark and Altafini (2018) for the case of continuous time systems.

4.2 Similarity to past metrics

In Hamdan and Nayfeh (1989), the cosine of the angle between an input column and the modal vector is defined as a metric. Our metric is strongly correlated with that metric, but our metric has the added benefit of allowing comparison across modes. Specifically, while considering the modal controllability of the i^{th} mode from the j^{th} input column, if we consider the metric in Hamdan and Nayfeh (1989), it is given by $\cos(\theta)_{ij}$. Whereas our metric for controlling the i^{th} mode from the j^{th} input is :

$$M_i = \frac{\|\mathbf{b}_j\|^2 \cos^2(\theta)_{ij}}{1-\lambda_i^2},$$

where $\mathbf{b}_j$ is the j^{th} input, and λ_i is the eigenvalue associated with the i^{th} mode.

4.3 Similarity to eigenvector centrality

We extend our ideas to linear network systems, by studying the relationship between our metric and the eigenvector centrality measure. In network science, there are several centrality measures (Bullo (2022)) such as degree, pagerank, and eigenvector centrality among others. Specifically we focus on the eigenvector centrality measure. The eigenvector centrality measure computes the influence of a node in a network. In practice, this centrality measure can be represented by the entries of the dominant eigenvector of the adjacency matrix. Specifically, if the dominant

eigenvalue is λ, and the associated eigenvector is $\mathbf{v}$, the centrality measure of the i^{th} node is given by $\mathbf{v}_i$. In the case of linear network systems, each column of the B matrix is a canonical vector. Therefore, our metric for infinite horizon, for the control of the dominant mode from the i^{th} input reduces to

$$M_{dom} = \frac{\mathbf{v}_i^2}{1-\lambda^2}. \tag{4}$$

We show that this is true for a 100 node Erdos-Renyi random network (Erdos et al. (1960)), by plotting M_{dom} against the eigenvector centrality in Fig. 1.

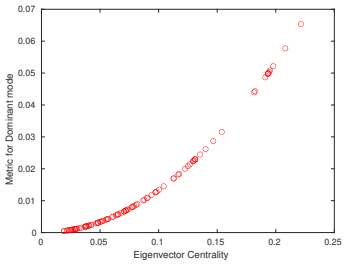


Fig. 1. The eigenvector centrality is related to the metric for the dominant mode for the i^{th} node as input as per Equation (4).

4.4 Modularity of the metrics

We now discuss the modularity properties of our proposed metrics.

Theorem 4.1. *For a given mode i, associated with the eigenvector $\mathbf{v}_i$, the function*

$$f(\mathcal{K}) = \mathbf{v}_i B_{\mathcal{K}} B_{\mathcal{K}}' \mathbf{v}_i \frac{1}{1-\lambda_i^2}.$$

is modular in $\mathcal{K}$.

Proof. Since each column in $B_{\mathcal{K}}$ is a canonical vector of dimension n, $B_{\mathcal{K}} B_{\mathcal{K}}' = \sum_{j\in\mathcal{K}} \mathbf{b}_j \mathbf{b}_j'$, our function reduces to

$$f(\mathcal{K}) = \sum_{j\in\mathcal{K}} \mathbf{v}_i' \mathbf{b}_j \mathbf{b}_j' \mathbf{v}_i \frac{1}{1-\lambda_i^2}.$$

Using Definition 2.1, the function is modular in $\mathcal{K}$. □

4.5 The case of Complex Modes

In general the system matrix A may have both real and complex eigenvalues and eigenvectors. Here, we briefly discuss the equivalent theory for complex modes. Using a similar line of arguments, and replacing the transpose operation with the conjugate transpose operation, the results and inequalities can be extended to modes with both real and imaginary components. For D-LTI systems,

$$M_i^d(t) = \left(\mathbf{v}_i^* B B^* \mathbf{v}_i \frac{1-|\lambda_i|^{2(t+1)}}{1-|\lambda_i^2|} \right).$$

5. CONCLUSION

We have considered the problem of modal controllability for networked systems. We propose an energy related metrics, for discrete LTI systems, which serves as a lower bound for the control effort required to move along a given mode. We analyze the metric by studying its relationship with existing metrics and other centrality measures, showing that it is closely related to the eigenvector centrality. We also formulate two problems regarding the optimization of the metrics, in an effort to minimize the control effort required for the control of a network along particular modes. Our approach to solve these problems are illustrated using a numerical example using the topology of a power grid.

REFERENCES

Bullo, F. (2022). *Lectures on Network Systems*. Kindle Direct Publishing, 1.6 edition. URL `http://motion.me.ucsb.edu/book-lns`.

Cortesi, F.L., Summers, T.H., and Lygeros, J. (2014). Submodularity of Energy Related Controllability Metrics. *arXiv e-prints*, arXiv:1403.6351.

Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1), 17–60.

Gu, S., Pasqualetti, F., Cieslak, M., Telesford, Q.K., Yu, A.B., Kahn, A.E., Medaglia, J.D., Vettel, J.M., Miller, M.B., Grafton, S.T., et al. (2015). Controllability of structural brain networks. *Nature communications*, 6(1), 1–10.

Hamdan, A. and Nayfeh, A. (1989). Measures of modal controllability and observability for first-and second-order linear systems. *Journal of guidance, control, and dynamics*, 12(3), 421–428.

Hespanha, J.P. (2018). *Linear systems theory*. Princeton university press.

Horn, R.A. and Johnson, C.R. (2012). *Matrix analysis*. Cambridge university press.

Lindmark, G. and Altafini, C. (2018). Minimum energy control for complex networks. *Scientific reports*, 8(1), 1–14.

Lovász, L. (1983). Submodular functions and convexity. In *Mathematical programming the state of the art*, 235–257. Springer.

Müller, P. and Weber, H. (1972). Analysis and optimization of certain qualities of controllability and observability for linear dynamical systems. *Automatica*, 8(3), 237–246.

Pasqualetti, F., Zampieri, S., and Bullo, F. (2014). Controllability metrics, limitations and algorithms for complex networks. *IEEE Transactions on Control of Network Systems*, 1(1), 40–52.

Summers, T.H., Cortesi, F.L., and Lygeros, J. (2016). On submodularity and controllability in complex dynamical networks. *IEEE Transactions on Control of Network Systems*, 3(1), 91–101. doi: 10.1109/TCNS.2015.2453711.

Summers, T.H. and Lygeros, J. (2014). Optimal sensor and actuator placement in complex dynamical networks. *IFAC Proceedings Volumes*, 47(3), 3784–3789.

Yan, G., Ren, J., Lai, Y.C., Lai, C.H., and Li, B. (2012). Controlling complex networks: How much energy is needed? *Physical review letters*, 108(21), 218703.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Noncommutative Optimal Polynomial Approximants

Palak Arora * **Meric Augat** ** **Michael Jury** ***
Meredith Sargent ****

* *University of Florida, Gainesville, FL, USA 32117.*
** *Washington University in St. Louis, St. Louis, MO 63110.*
*** *University of Florida, Gainesville, FL, USA 32117.*
**** *University of Manitoba, Winnipeg, MB, Canada R3T 2N2*

Abstract: The study of Optimal Polynomial Approximants (OPAs) in weighted Dirichlet-type spaces has seen a great deal of success in the past decade. In more than one variable, not as much is known, and the failure of the famous *Shanks Conjecture* shows that there may be issues with the typical Drury-Arveson approach.
A potential remediation is to recast the multivariable into the noncommutative setting where the deficiencies of the Drury-Arveson space are not found. This paper outlines the introductory ideas behind noncommutative Optimal Polynomial Approximants, as well as a couple of tangible conjectures and potential approaches inspired by classical techniques in the freely noncommutative setting.

Keywords: Optimal Polynomial Approximants, Noncommutative Polynomials, Realizations, Row Ball, Shanks Conjecture, Fock Space

AMS Subject Classification: 47A56, 30H20

1. INTRODUCTION

In the last decade, there has been considerable mathematical literature produced about optimal polynomial approximants in connection with Hilbert spaces of analytic functions (of one or several variables) – Bénéteau and Centner (2021); Bénéteau et al. (2013, 2018, 2016a); Sargent and Sola (2020a,b) to name a few. Recall the Hardy space H^2 is the space of functions $f(z) = \sum_{k=0}^{\infty} a_k z^k$ analytic in $\mathbb{D}$, the unit disk, such that $\|f\|^2 = \sum_{k=0}^{\infty} |a_k|^2 < \infty$. We define an inner-product for H^2 by $\langle f, g\rangle := \sum_{k=0}^{\infty} a_k \overline{b_k}$, where $g(z) = \sum_k b_k z^k$. Given a nonzero function $f \in H^2$ and $n \in \mathbb{N}$, we define the n^{th} **optimal polynomial approximant** (OPA) of $1/f$ in H^2 to be the polynomial q_n that minimizes $\|pf - 1\|$ among all polynomials p of degree at most n. While only presented here in terms of the Hardy space, OPAs have undergone considerable investigation in terms of Dirichlet-type Hilbert spaces.

Another related and highly studied idea are cyclic functions Bénéteau et al. (2016b, 2013, 2015). A function $f \in H^2$ is **cyclic** if $\{pf : p \in \mathbb{C}[z]\}$ is dense in H^2. It turns out that a function f is cyclic if and only if the OPAs q_n of f satisfy $\|q_n f - 1\| \to 0$ as $n \to \infty$. Thus, we can study cyclic functions via OPAs and vice versa.

A strange occurrence in commutative functional analysis in several variables is that while analogues of many classical Hardy space results can be found, their proofs often balloon in complexity or in the use of novel proof techniques. One explanation can be found in Jury and Martin (2020): with the Hardy space in one variable, multiplication by z is the *shift operator* and is an isometry. Moreover, every isometry is unitarily equivalent to a direct sum of shifts and a unitary operator. However, the natural multivariable generalization of the Hardy space is the Drury-Arveson space, where the appropriate analogue of the shift is the Arveson d-shift $S := (S_1, \ldots, S_d)$ with $S_i f = z_i f$. Notably, the Arveson d-shift is no longer a (row) isometry, but rather a (row) partial isometry, which is the source of many defects moving from one to several variables. In Jury and Martin (2020), the authors correctly identify the spiritual successor of the Hardy space shift in several variables: the left free shift on the full Fock space. Their choice is justified by the strength of their conclusions: multivariable generalizations of Hardy space theorems that also persist under compression to commuting variables.

Recall that $\mathbb{C}\langle x\rangle$ is the free algebra in d freely noncommuting indeterminates. We make $\mathbb{C}\langle x\rangle$ into a pre-Hilbert space by choosing the free monoid $\langle x\rangle$ to be an orthonormal basis. That is, if $w_1 = x_{i_1} \ldots x_{i_n}$ and $w_2 = x_{j_1} \ldots x_{j_m}$, then $\langle w_1, w_2\rangle$ equals 1 if $n = m$ and $i_k = j_k$, and it equals 0 otherwise. The **full Fock space** in d-letters $\mathcal{F}_d$ is the completion of $\mathbb{C}\langle x\rangle$ with respect to this inner product. The **left creation operators** $L_1, \ldots, L_d$ are natural analogues of the Hardy space shift in the sense that if $f \in \mathcal{F}_d$, then $L_j f = x_j f$ – naturally, we can also define the *right creation operators* as $R_j f = f x_j$. The **left free shift** is the tuple $L = (L_1, \ldots, L_d)$; L is a row isometry, and given any row isometry, it is isomorphic to the direct sum of copies of L and a row unitary. Moreover, it has been well established in the noncommutative function theory literature Agler

and McCarthy (2015); Popescu (2006, 2010); Jury et al. (2021a,b) as well as in terms of noncommutative reproducing kernel Hilbert spaces Ball et al. (2016) that $\mathcal{F}_d$ is canonically isomorphic to the free Hardy space, $H^2(\mathbb{B}_{\mathbb{N}}^d)$ of noncommutative or free holomorphic functions on a certain noncommutative multivariable open unit ball, $\mathbb{B}_{\mathbb{N}}^d$.

With respect to generalizing OPAs to several variables, as we see from above, the full Fock space is a more faithful multivariable generalization of the Hardy space. Moreover, the recent papers Jury et al. (2021a,b) fully characterize cyclic functions in $\mathcal{F}_d$. A function $f \in \mathcal{F}_d$ is **cyclic (for the left free shift)** if its left polynomial multiples have dense range in $\mathcal{F}_d$ if and only if $f(X)$ is nonsingular for all X in the row ball. That is, the function f has no singularities in the unit ball, a condition that is reminiscent of the situation in the classical Hardy space.

We now arrive at the notion of a noncommutative optimal approximant: given nonzero $f \in \mathcal{F}_d$ and $n \in \mathbb{N}$, the n^{th} left nc optimal polynomial approximant (nc OPA) of $1/f$ is the free polynomial p_n that minimizes the norm $\|pf - 1\|_{\mathcal{F}_d}$ among all free polynomials p of degree at most n. Note that in every practical sense, the theory of left nc OPAs is identical to the theory of right nc OPAs. Taking advantage of the theory of reproducing kernels for NC functions, we arrive at the following natural generalization of a classical result.

Conjecture 1. Suppose f is a free polynomial that is cyclic for the left free shift. If p_n is the (left) nc OPA of $1/f$ and $p_n(\Lambda) = 0$ for some matrix tuple $\Lambda = (\Lambda_1, \ldots, \Lambda_d)$, then $\|\Lambda_1 \Lambda_1^* + \cdots + \Lambda_d \Lambda_d^*\| > 1$.

At the time of writing this abstract, the conjecture above remains open. However, the novel techniques used in investigating the conjecture have already borne fruit in the commutative setting:

Theorem 2. Suppose f is a commutative polynomial that is cyclic for the Drury-Arveson d-shift. If q_n is the n^{th} OPA of $1/f$ in the Drury-Arveson space, then q_n has no zeros in the row ball.

That is, the row ball version of the *Shanks Conjecture* holds.

Noncommutative arguments can often be obtained from single variable proofs by generalizing the proof with "one hand tied behind your back," i.e. making no use of commutativity. These nc arguments often generalize readily to several variables. One distinct occasion where this fails, is when the commutative argument takes advantage of factoring, since in the noncommutative setting factoring is often not possible e.g. $x_1x_2x_1 + x_2$ is irreducible when x_1 and x_2 do not commute. However, the resolution to this problem is something called *stable associativity* – see Cohn (2006). Two free polynomials f and g are **stably associated**, if there exist $m \in \mathbb{Z}^+$ and $P, Q \in \mathrm{GL}_{m+1}(\mathbb{C}\langle x \rangle)$ such that

$$P(f \oplus I_m) = (g \oplus I_m)Q.$$

For example, xy and yx are stably associated. While stable associativity seems like an inconvenience, it does have advantages: every atomic (cannot be decomposed as a nontrivial product of free polynomials) free polynomial f with $f(0) = I$ is stably associated to an irreducible monic linear pencil $\mathcal{L}$. This is incredibly appealing, since monic linear pencils are remarkably well-studied Klep and Volčič (2017); Klep et al. (2016); Helton et al. (2017). If we are able to extend our notions of nc OPAs to matrix valued functions, then every problem can be stated in terms of an affine-linear (matrix-valued) polynomial. Thus, we are led to a pair of questions.

Problem 1. What properties of nc OPAs are preserved under stable associativity?

Problem 2. Can nc OPAs be generalized to matrix-valued free polynomials? If so, what properties are preserved under stable associativity?

REFERENCES

Agler, J. and McCarthy, J.E. (2015). Global holomorphic functions in several noncommuting variables. *Canadian Journal of Mathematics*, 67(2), 241–285.

Ball, J.A., Marx, G., and Vinnikov, V. (2016). Noncommutative reproducing kernel hilbert spaces. *Journal of Functional Analysis*, 271.

Bénéteau, C. and Centner, R. (2021). A survey of optimal polynomial approximants, applications to digital filter design, and related open problems. *Complex Analysis and its Synergies*, 7.

Bénéteau, C., Condori, A., Liaw, C., Seco, D., , and Sola, A. (2013). Cyclicity in dirichlet-type spaces and extremal polynomials ii: functions on the bidisk. *Pacific Journal of Mathematics*, 276.

Bénéteau, C., Condori, A.A., Liaw, C., Seco, D., , and Sola, A.A. (2015). Cyclicity in dirichlet-type spaces and extremal polynomials. *Journal d'Analyse Mathématique*, 126.

Bénéteau, C., Fleeman, M.C., Khavinson, D.S., Seco, D., and Sola., A.A. (2018). Remarks on inner functions and optimal approximants. *Canadian Mathematical Bulletin*.

Bénéteau, C., Khavinson, D., Liaw, C., Seco, D., and Sola., A.A. (2016a). Orthogonal polynomials, reproducing kernels, and zeros of optimal approximants. *Transactions of the American Mathematical Society*, 94.

Bénéteau, C., Knese, G., Kosiński, Ł., Liaw, C., Seco, D., and Sola., A.A. (2016b). Cyclic polynomials in two variables. *Transactions of the American Mathematical Society*, 368.

Cohn, P.M. (2006). *Free Ideal Rings and Localization in General Rings*. New Mathematical Monographs. Cambridge University Press. doi: 10.1017/CBO9780511542794.

Helton, J.W., Klep, I., and McCullough, S. (2017). The tracial Hahn–Banach theorem, polar duals, matrix convex sets, and projections of free spectrahedra. *Journal of the European Mathematical Society*, 19(6), 1845–1897.

Jury, M., Martin, R., and Shamovich, E. (2021a). Noncommutative rational functions in the full fock space. *Transactions of the American Mathematical Society*.

Jury, M.T. and Martin, R.T.W. (2020). Column extreme multipliers of the free hardy space. *Journal of the London Mathematical Society*, 101.

Jury, M.T., Martin, R.T., and Shamovich, E. (2021b). Blaschke–singular–outer factorization of free noncommutative functions. *Advances in Mathematics*, 384.

Klep, I. and Volčič, J. (2017). Free loci of matrix pencils and domains of noncommutative rational functions. *Commentarii Mathematici Helvetici*, 92(1), 105–130.

Klep, I., Vinnikov, V., and Volčič, J. (2016). Multipartite rational functions. URL `https://arxiv.org/abs/1509.03316`.

Popescu, G. (2006). Free holomorphic functions on the unit ball of $\mathcal{B}(\mathcal{H})^n$. *Journal of Functional Analysis*, 241(1), 268–333.

Popescu, G. (2010). Free holomorphic automorphisms of the unit ball of $\mathcal{B}(\mathcal{H})^n$. *Journal für die reine und angewandte Mathematik*, 2010(638), 119–168.

Sargent, M. and Sola, A.A. (2020a). Optimal approximants and orthogonal polynomials in several variables. *Canadian Journal of Mathematics*.

Sargent, M. and Sola, A.A. (2020b). Optimal approximants and orthogonal polynomials in several variables ii: families of polynomials in the unit ball.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

On the Implementation and Adaptation of a Class of Internal Models ⋆

Patrizio Colaneri * **Gian Paolo Incremona** * **Lorenzo Marconi** **
Leonid Mirkin ***

* *DEIB, Politecnico di Milano, 20133 Milan, Italy*
(e-mails: patrizio.colaneri@polimi.it & gianpaolo.incremona@polimi.it)
** *CASY-DEI, University of Bologna, Bologna 40123, Italy*
(e-mail: lorenzo.marconi@unibo.it)
*** *Faculty of Mechanical Eng., Technion—IIT, Haifa 3200003, Israel*
(e-mail: mirkin@technion.ac.il)

Abstract: A recent result in (Incremona et al., 2022) put forward an architecture of internal model based controllers, in which the stabilizer can be fully separated from the internal model. In this paper we propose a parametrized implementation of this controller, which isolates a parameter shaping properties of the exosystem. We show that with this implementation the closed-loop dynamics have an affine dependence on the parameter. As such, the closed-loop system remains stable even under arbitrary variations of the parameter, as long as it remains bounded. We demonstrate that this property is beneficial for adding an adaptation mechanism to adjust parameters of the internal model.

Keywords: Regulator problem, Internal Model Principle, adaptive control.

1. INTRODUCTION

Consider the plant

$$\dot{x}(t) = Ax(t) + Bd(t) + Bu(t), \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is its measurable state, $u(t) \in \mathbb{R}^m$ is a control input, and $d(t) \in \mathbb{R}^m$ is a load (matched) disturbance. We assume that the disturbance is generated by the *exosystem*

$$\begin{cases} \dot{x}_{\rm ex}(t) = A_{\rm ex}x_{\rm ex}(t), \quad x_{\rm ex}(0) = x_{\rm ex,0} \\ d(t) = C_{\rm ex}x_{\rm ex}(t) \end{cases} \tag{2}$$

for known $A_{\rm ex}$, assumed to possess no open left half-place eigenvalues, and $C_{\rm ex}$ and an unknown initial condition $x_{\rm ex,0}$. The control goal is to have an asymptotic rejection of every disturbances of this class, i.e. $\lim_{t\to\infty} x(t) = 0$. This is a particular case of the classical regulator problem, see (Saberi et al., 2000; Isidori, 2017) and the references therein.

The regulator problem can be solved via the use of the internal model principle of Francis and Wonham (1975). A conventional modus operandi is to transplant a model of exosystem (2), known as an *internal model*, into the feedback loop and then stabilize the *augmented system*, comprising both the plant and the internal model. The internal model needs to have the eigenvalues of $A_{\rm ex}$ as its poles, but the choice of its other properties is more flexible and can be adjusted to a concrete architecture. This procedure is well understood in the linear case, even for a more general setup than that described above.

Arguably, one of main shortcomings of the design procedure outlined above is the need to design the stabilizing part of the controller for augmented dynamics. This is not a serious issue if (2) is low dimensional, but might become a problem for more complex exosystems, like those used in repetitive control (Longman, 2010). Moreover, changes in parameters of the internal model necessitate a redesign of the stabilizer. This could substantially complicate the employment of various adaptation methods and deter from the use of high-order models, which could be useful in reducing the sensitivity to small deviations of d from its modeled version (Singhose, 2009).

An alternative approach to incorporate internal models into the feedback loop was put forward in (Incremona et al., 2022). Its essence is the use of an "internal model compensation" element in addition to the internal model itself, inspired by the delay compensation idea in (Mirkin, 2020) for repetitive control. This addition enables a complete separation of the stabilizer, which should be designed only for the unaltered plant (1), from the internal model.

Circumventing the need to deal with augmented dynamics simplifies the design of the stabilizer, it can be a static state feedback in the studied case. Yet the dependence of the resulting controller on parameters of the exosystem is still complex, which hampers adjusting those parameters. The goal of this paper is to propose an alternative implementation of the controller of (Incremona et al., 2022), in which parameters of the exosystem that shape its properties, like oscillation frequencies, can be isolated. The proposed architecture simplifies the controller implementation and, more importantly, substantially simplifies the dependence of the closed-loop dynamics on those parameters. Specifically, we show that the dependence of the closed-loop system on those parameters is affine. As a result, the closed-loop stability is maintained even if parameters of the internal model are tuned, provided they remain bounded.

This result paves the way to incorporating adaptation mechanisms, whose purpose is to tune the model to uncertain / changing exosystems. We demonstrate the usefulness of the proposed

⋆ Supported by the Israel Science Foundation (grant no. 3177/21) and Sakranut Graydah.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

architecture by presenting an adaptation mechanism, which ensures the global convergence and asymptotic attenuation of a given class of disturbances under arbitrary unknown parameters of the disturbance model.

Notation Given a full column rank matrix $M \in \mathbb{R}^{n\times m}$, its left inverse is denoted as $M^{\#} \in \mathbb{R}^{m\times n}$, so $M^{\#}M = I$. We use the compact notation

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] := D + C(sI - A)^{-1}B$$

for transfer functions in terms of their state-space realizations. The lower linear-fractional transformation

$$\mathcal{F}_{\mathrm{l}}\left(\begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}, \Omega\right) := \Phi_{11} + \Phi_{12}\Omega(I - \Phi_{22}\Omega)^{-1}\Phi_{21},$$

see (Zhou et al., 1996, Ch. 10) for its properties.

2. PARAMETRIZATION OF THE INTERNAL MODEL

We start with fixing the structure of the control law in the form

$$u = M(v + R_{\mathrm{stab}}x). \tag{3}$$

Here M is an internal model, whose poles coincide with those of exosystem (2), R_{stab} is a stabilizer, whose purpose is naturally to stabilize the closed-loop system, and v is a signal that can be used to introduce a reference signal. We assume that

$\boldsymbol{\mathcal{A}_1}$: $M(\infty) = I$ and $M^{-1} \in H_\infty$,

which is required by the design procedure of (Incremona et al., 2022) and entails no loss of generality (it does not affect poles).

A key component of the proposed approach is a special parametrization of the internal model in the form

$$\begin{aligned} M^{-1}(s) &= \left[\begin{array}{c|c} A_0 & B_0 \\ \hline C_{01} & I \end{array}\right] + \Theta \left[\begin{array}{c|c} A_0 & B_0 \\ \hline C_{02} & 0 \end{array}\right] \\ &= \mathcal{F}_{\mathrm{l}}\left(\left[\begin{array}{c|cc} A_0 & B_0 & 0 \\ \hline C_{01} & I & I \\ C_{02} & 0 & 0 \end{array}\right], \Theta\right) \end{aligned} \tag{4}$$

for a real parameter Θ and a Hurwitz A_0. In the SISO case this implies that we assume that the denominator of $M(s)$ is an affine function of parameters and its numerator is Hurwitz and does not depend on these parameters. In general, Θ shapes the spectrum of A_{ex} in (2). To see that, note that

$$\begin{aligned} M(s) &= \mathcal{F}_{\mathrm{l}}\left(\left[\begin{array}{c|cc} A_0 - B_0C_{01} & B_0 & -B_0 \\ \hline -C_{01} & I & -I \\ C_{02} & 0 & 0 \end{array}\right], \Theta\right) \\ &= \left[\begin{array}{c|c} A_0 - B_0(C_{01} + \Theta C_{02}) & B_0 \\ \hline -C_{01} - \Theta C_{02} & I \end{array}\right], \end{aligned}$$

which follows from (Zhou et al., 1996, Lem. 10.3). This naturally leads to the assumption that

$\boldsymbol{\mathcal{A}_2}$: $A_{\mathrm{ex}} = A_0 - B_0(C_{01} + \Theta C_{02})$ and $C_{\mathrm{ex}} = -C_{01} - \Theta C_{02}$,

which shall guarantee that the internal model solves the regulator problem.

Representation (4) is related to the representation of an exosystem in (Nikiforov, 1997). A difference is that we *assume* this form of the internal model, whereas it was derived in (Nikiforov, 1997, Lem. 3.1) as an asymptotic model for a general SISO exosystem.

To provide a flavor of this structure, consider a couple of simple examples. First, if $d(t) = a\sin(\omega t + \phi)$ with a known frequency $\omega > 0$ and unknown amplitude a and phase ϕ, then a general internal model satisfying $\boldsymbol{\mathcal{A}_1}$ is $M(s) = \phi(s)/(s^2 + \omega^2)$ for an arbitrary second-order Hurwitz and monic polynomial $\phi(s)$. In this case

$$M^{-1}(s) = \frac{s^2}{\phi(s)} + \frac{\Theta}{\phi(s)} = \mathcal{F}_{\mathrm{l}}\left(\begin{bmatrix} s^2/\phi(s) & 1 \\ 1/\phi(s) & 0 \end{bmatrix}, \Theta\right) \tag{5}$$

under $\Theta = \omega^2$.

The case of two harmonics, say with frequencies $\omega_1 > 0$ and $\omega_2 > 0$, is perhaps less obvious. For example, the sum of two harmonic oscillators as above does not fit into (4), because the parameters (frequencies) affect its numerator as well. But it is not hard to see that $M(s) = \phi(s)/((s^2 + \omega_1^2)(s^2 + \omega_2^2))$ for an arbitrary fourth-order Hurwitz and monic $\phi(s)$ is an admissible choice. It corresponds to

$$\begin{aligned} M^{-1}(s) &= \frac{s^4}{\phi(s)} + \frac{(\omega_1^2 + \omega_2^2)s^2 + \omega_1^2\omega_2^2}{\phi(s)} \\ &= \mathcal{F}_{\mathrm{l}}\left(\left[\begin{array}{c:c} s^4/\phi(s) & 1 \\ \hdashline s^2/\phi(s) & 0 \\ 1/\phi(s) & 0 \end{array}\right], \Theta\right) \end{aligned}$$

under $\Theta = \begin{bmatrix} \omega_1^2 + \omega_2^2 & \omega_1^2\omega_2^2 \end{bmatrix}$, which is in a bijective relation with ω_1^2 and ω_2^2.

3. MAIN RESULTS

Given any internal model M satisfying $\boldsymbol{\mathcal{A}_1}$, the procedure of (Incremona et al., 2022) is to select the stabilizer of the form

$$R_{\mathrm{stab}} = \tilde{R}_{\mathrm{stab}} - \Upsilon,$$

where the "internal model compensation" Υ is any *stable* system such that

$$\Upsilon(s)(sI - A)^{-1}B = I - M^{-1}(s)$$

and $\tilde{R}_{\mathrm{stab}}$ is any controller stabilizing the (non-augmented) plant (1). Possible choices are

$$\Upsilon(s) = (I - M^{-1}(s))(B^{\#}s - B^{\#}A), \tag{6}$$

where $B^{\#}$ is a left inverse of B, and

$$\tilde{R}_{\mathrm{stab}} = K,$$

where $K \in \mathbb{R}^{m\times n}$ is such that $A + BK$ is Hurwitz (the full column rank of B and stabilizability of (A, B) are naturally assumed). This yields the following form of the control law (3):

$$u = M(v + (K - \Upsilon)x). \tag{7}$$

Controller (7) involves two dynamic elements, the internal model M and the internal model compensator Υ. They both have a complex dependence on the model parameter Θ, which hampers studying the effect of this parameter on the closed-loop dynamics. So our first goal is to present (7) in an alternative form, in which the parameter Θ is isolated. This is done by the result below, whose proof is omitted.

Proposition 1. The controller $R : (v, x) \mapsto u$ in (7) can be implemented as $R(s) = \mathcal{F}_{\mathrm{l}}(\Psi, \Theta)$, see Fig. 1(a), where

$$\Psi(s) = \left[\begin{array}{c|c:c:c} A_0 - B_0C_{01} & B_0 & B_K & -B_0 \\ \hline -C_{01} & I & K + C_{01}B_0B^{\#} & -I \\ \hdashline C_{02} & 0 & -C_{02}B_0B^{\#} & 0 \end{array}\right] \tag{8}$$

and $B_K := B_0B^{\#}(A + BK) - (A_0 - B_0C_{01})B_0B^{\#}$.

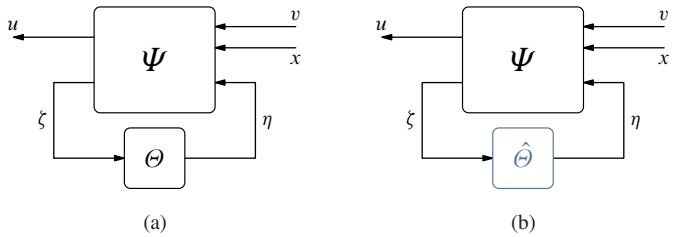


Fig. 1. Implementation of the controller $R : (v, x) \mapsto u$

Plugging this controller into the plant dynamics (1), we have the following expression for the resulting closed-loop system (the proof is also omitted).

Proposition 2. The closed-loop system $T : (v, d) \mapsto (x, u)$ is

$$T = T_v \begin{bmatrix} I & M^{-1} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \tag{9}$$

where

$$T_v(s) = \left[\begin{array}{c|c} A + BK & B \\ \hline I & 0 \\ K & I \end{array}\right].$$

Some remarks are in order.

- The effect of the signal v on the closed-loop system, which is $T_v : v \mapsto (x, u)$, is *independent* of the internal model and its shaping parameter Θ. Thus, consistently with the result of (Incremona et al., 2022), we can design v as if no internal model was present in the loop. This could simplify tracking design, as well as facilitates designing an external loop, which may be helpful to handle an unmodeled part of d.
- Straightforward algebra, together with $\boldsymbol{\mathcal{A}_2}$, yields that the signal $d_0 = M^{-1}d$ satisfies

$$\begin{cases} \dot{x}_0(t) = A_0 x_0(t), \quad x_0(0) = x_{\text{ex},0} \\ d_0(t) = C_{\text{ex}} x_0(t). \end{cases} \tag{10}$$

  This is an exponentially decaying function. The stability of T_v implies then that the effect of the disturbance, at least of its part modeled by (2), decays in steady state. Thus, if $\boldsymbol{\mathcal{A}_2}$ holds, i.e. if the internal model M agrees with the exosystem, then the regulator problem is solved in the sense that

$$\lim_{t\to\infty} x(t) = 0 \quad \text{and} \quad \lim_{t\to\infty} (u(t) + d(t)) = 0.$$

4. ADAPTATION MECHANISM FOR UNKNOWN Θ

Advantages of the proposed architecture are even sharper in the case of a varying Θ. Changing one or several parameters in conventional internal model configurations might result in a need to redesign the stabilizer and could thus substantially complicate the analysis. Yet changing Θ in Fig. 1(a) results in the very same closed-loop system as in (9), just with a varying Θ affecting affinely M^{-1} in (4). And the closed-loop system is stable as long as Θ remains bounded. This property paves the way to the design of globally converging adaptation algorithms, in which Θ in the internal model is adjusted to match unknown and possibly varying parameters of the exosystem.

Specifically, assume that Θ is constant, but unknown. Denote by $\hat{\Theta}$ its estimate, to be defined later on. In this case instead of the controller in Fig. 1(a) we implement that in Fig. 1(b), i.e. just replace Θ with its estimate. The following result can be formulated.

Proposition 3. Introduce the signal $e(t) := x(t) - x_v(t)$, where

$$\dot{x}_v(t) = (A + BK)x_v(t) + Bv(t),$$

and let $P = P' > 0$ satisfy $(A + BK)'P + P(A + BK) < 0$ (it exists because $A + BK$ is Hurwitz). The adaptation law

$$\dot{\hat{\Theta}}(t) = B'Pe(t)\zeta'(t), \quad \hat{\Theta}(0) = \hat{\Theta}_0 \tag{11}$$

where ζ is the second output of Φ in Fig. 1(b), guarantees the global boundedness of all signals in the system and the asymptotic regulation in the sense $\lim_{t\to\infty} e(t) = 0$.

Proof (outline). A key observation, which can be shown by straightforward algebra, is that the closed-loop plant satisfies

$$\dot{x}(t) = (A + BK)x(t) + Bv(t) + Bd_0(t) + B\tilde{\Theta}(t)\zeta(t),$$

where $\tilde{\Theta}(t) := \Theta - \hat{\Theta}(t)$ is the parameter mismatch and d_0 satisfies (10). Hence,

$$\dot{e}(t) = (A + BK)e(t) + Bd_0(t) + B\tilde{\Theta}(t)\zeta(t)$$

is independent of v. Because d_0 is exponentially decaying and independent of u, it can be excluded from the stability analysis. Consider the Lyapunov candidate

$$V(t) = e'(t)Pe(t) + \operatorname{tr}(\tilde{\Theta}(t)'\tilde{\Theta}(t))$$

for which, taking into account that $\dot{\tilde{\Theta}} = -\dot{\hat{\Theta}}$,

$$\dot{V}(t) = -e'(t)Qe(t) - 2\operatorname{tr}(\tilde{\Theta}'(t)(\dot{\hat{\Theta}}(t) - B'Pe(t)\zeta'(t))).$$

where $Q = -(A+BK)'P - P(A+BK) > 0$. Hence, (11) yields $\dot{V}(t) = -e'(t)Qe(t) \leq 0$. The result follows then by standard adaptive control arguments, like those in (Lavretsky and Wise, 2013, Ch. 9). □

4.1 Illustrative example

Consider the plant with

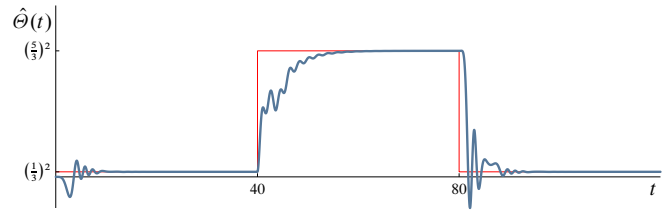
$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and $d(t) = 0.5\sin(\omega t)$. We model this disturbance by (5) with the Butterworth $\phi(s) = s^2 + \sqrt{2}s + 1$. The state-feedback gain

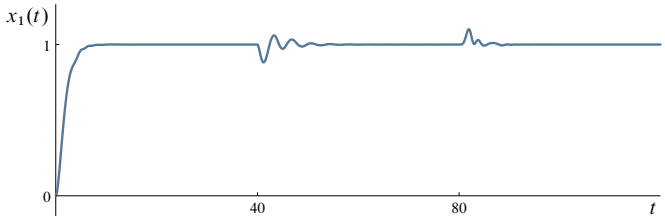
$$K = \begin{bmatrix} -1 & -2 \end{bmatrix},$$

assigning both eigenvalues of $A + BK$ to -1. We choose the the unit step $v(t)$. Finally, in the adaptation law (11) we select $P = 75\left[\begin{smallmatrix} 3 & 1 \\ 1 & 1 \end{smallmatrix}\right]$, which is the solution to the Lyapunov equation $(A + BK)'P + P(A + BK) = -150I < 0$.

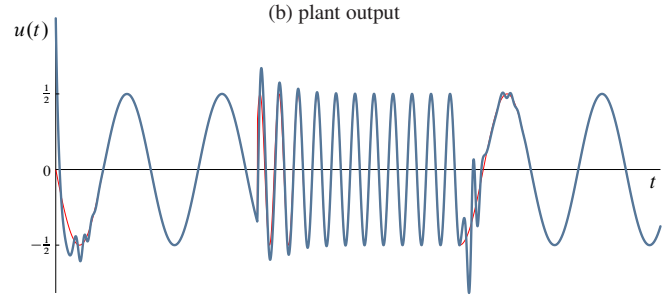
We simulate the resulted closed-loop system for ω switching between $1/3$ and $5/3$ every 40 time units. The estimated parameter $\hat{\Theta}(t) = \hat{\omega}^2(t)$ is shown in Fig. 2(a). Note that it might take (nonphysical) negative values during transients. To prevent this, the adaptation law (11) has to be modified, for example in line with the discussion in (Goodwin and Mayne, 1987). But it eventually always converges to the true value of Θ in this example. The first component of the state, $x_1(t)$, is presented in Fig. 2(b). Its steady state $\lim_{t\to\infty} x_1(t) = 1$ despite harmonic disturbances, which is exactly why we introduced the internal model to the controller. The control signal, shown in Fig. 2(c), expectably converges to $-d(t)$, which is required to cancel the effect of the load disturbance.



(a) estimated disturbance parameter (actual Θ is shown by the thin red line)



(b) plant output



(c) control signal (the additive inverse of $d(t)$ is shown by the thin red line)

Fig. 2. Simulation results

REFERENCES

Francis, B.A. and Wonham, W.M. (1975). The internal model principle for linear multivariable regulators. *Appl. Math. Opt.*, 2(2), 170–412.

Goodwin, G.C. and Mayne, D.Q. (1987). A parameter estimation perspective of continuous time model reference adaptive control. *Automatica*, 23(1), 57–70.

Incremona, G.P., Mirkin, L., and Colaneri, P. (2022). Integral sliding-mode control with internal model: A separation. *IEEE Control Syst. Lett.*, 6, 446–451.

Isidori, A. (2017). *Lectures in Feedback Design for Multivariable Systems*. Springer-Verlag, Cham, CH.

Lavretsky, E. and Wise, K.A. (2013). *Robust and Adaptive Control with Aerospace Applications*. Springer-Verlag, London.

Longman, R.W. (2010). On the theory and design of linear repetitive control systems. *European J. Control*, 16(5), 447–496.

Mirkin, L. (2020). On dead-time compensation in repetitive control. *IEEE Control Syst. Lett.*, 4(4), 791–796.

Nikiforov, V.O. (1997). Adaptive servomechanism controller with an implicit reference model. *Int. J. Control*, 68(2), 277–286.

Saberi, A., Stoorvogel, A.A., and Sannuti, P. (2000). *Control of Linear Systems with Regulation and Input Constraints*. Springer-Verlag, London, UK.

Singhose, W. (2009). Command shaping for flexible systems: A review of the first 50 years. *Int. J. Precis. Eng. Manuf.*, 10(4), 153–168.

Zhou, K., Doyle, J.C., and Glover, K. (1996). *Robust and Optimal Control*. Prentice-Hall, Englewood Cliffs, NJ.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Persistent Homology-based Resilience Enhancement

G. Revati [*], **S. Shadab** [*], **M. R. Mariya** [*], **A. Pandey** [**]
S. R. Wagh [*], **F. Kazi** [**], and **N. M. Singh** [*]

[*] *Control and Decision Research Centre (CDRC), EED, VJTI, Mumbai, India (e-mail: rgunjal_p21@ee.vjti.ac.in)*
[**] *Centre of Excellence-Complex Nonlinear Dynamical System (COE-CNDS), VJTI, Mumbai, India*

Abstract: The adverse effect of increasing penetration of distributed energy resources has resulted in increased vulnerabilities to resilience, defined as the ability of the grid to preserve the original properties under disruptive scenarios which were unforeseen in the traditional power grid. Hence it necessitates the development of accurate and reliable resilience metrics to have deeper insight under any disturbance resulting in modification of structural properties. Especially, considering the critical role of local structure and its inherent underlying geometry makes the impact analysis more challenging. In view of this, the proposed Persistent homology-based resilience enhancement (PHRE) technique utilizes the concept of Topological Data Analysis, particularly Betti numbers (identifying the most vulnerable buses) and persistent homology, extracting the longer-lasting topological features of the graphs through the network filtration at various spatial resolutions characterizing the structural functionality of the network. The proposed PHRE technique is validated using a benchmark system.

Keywords: Betti-number, Grid instability, Persistent homology, Topological analysis

1. INTRODUCTION

The resilience of the power grid Tajer et al. (2021) is referred to as the quantification of the ability of the grid network to maintain its functions under component failures from random errors or external causes. During the design of traditional power grids, the unforeseen vulnerabilities arising from the increasing penetration of renewable energy generations and distributed energy resources were not considered. Hence with the challenging assessment of grid organization, reliable resilience metrices are necessary for the energy systems under various disruptive scenarios, including component failure, various attacks, or natural disasters.

Understanding the structural properties of power networks under disruptive scenarios gives profound insights into their vulnerability. The most explored characteristics of power grid resilience are node degree distribution and mean degree, which are local properties Cuadra et al. (2015) investigated at the level of individual nodes and edges. Incorporating electrical engineering concepts like impedances and power flows Zhu et al. (2014) into the graph, particularly at the local level, then analyzing their impact on grid functionality is a further difficult endeavor. Recent studies Dey et al. (2017), Schultz et al. (2014) suggest that the power grid robustness is associated with its geometry through the network motifs, which are the multi-node sub-graph patterns. However, motifs are inefficient for power networks due to limited applicability to unweighted topological graphs that fail to reflect information about power grid functionality Sánchez-García et al. (2014). Hence the emerging technique called Topological Data Analysis (TDA) Patania et al. (2017), particularly, Persistent Homology (PH) Otter et al. (2017), applicable for weighted topological graphs is explored for the grid analysis since its flexibility of integrating with machine learning approaches.

The extended abstract proposes analysis of the power system resilience through a Persistent Homology-based Resilience Enhancement (PHRE) technique. Network geometry has characterised its structural functionality, and the system's ability to preserve the original network properties for a longer time during disruptive scenarios is enhanced through the PHRE technique for maintaining the resilience standard. The number of one-dimensional holes or the topological loops is identified as the resilience index. The healthier system reports a higher number of holes, and the number drops down drastically for the cascade failures. The PHRE approach analyses the cascade failures triggered by various causes such as voltage instability, overloading, etc., as it is purely based on the network topology and characterizes the functional information of the grid through its underlying geometry.

The extended abstract discusses the main idea of PHRE technique and the representative case study illustrating the implementation on the benchmark IEEE 30 bus system for validating the proposed approach.

2. TOPOLOGICAL DATA ANALYSIS (TDA) AND PERSISTENT HOMOLOGY (PH)

TDA combines various fields, including algebraic topology, data analysis, computational geometry, and statistics.

TDA aims to gain deeper insights into the qualitative characteristics of the data by utilizing the results obtained from its geometry and topology. It is appealing approach for the complex network functionality analysis through some crucial features unveiling the component interactions at multi-scale. These qualitative features are identified through one of the TDA techniques called PH.

PH computes topological features at various spatial resolutions. Consider a weighted graph $G = (V, E, w)$ with V, E, and w denoting the set of vertices, set of edges, and vector of edge weights, respectively. To find the PH of G, the hierarchically nested sequence of subgraphs $G_1 \subseteq G_2 \subseteq \cdots G_n$ with the increased threshold limits on edge weights $\upsilon_1 < \upsilon_2 < \cdots < \upsilon_n$ are obtained, and this stage is referred to as network filtration. Due to the computational efficiency, the Vietoris-Rips (VR) complex Zomorodian (2010) which considers the edge weights as a distance measure between the two nodes forming the given edge is used for network filtration. At a particular weight threshold υ_j the VR complex is defined as

$$\mathrm{VR}_j = \{\varrho \subset V | w_{pq} \leq \upsilon_j \forall p, q \in \varrho\} \tag{1}$$

where the edges with weights $w_{pq} \leq \upsilon_j$ are kept and the remaining edges are discarded to obtain the subgraph G_j with the adjacency matrix $A_{pq} = 1_{w_{pq} \leq \upsilon_j}$. VR_j will contain the k-node subsets of G_j where $k = 1, \cdots, K$ with the pairwise connection with an edge as simplices of dimensions $k-1$. These simplicial complexes aid in approximating the hidden geometric structures of the grid in a combinatorial way.

With VR filtration, $\mathrm{VR}_1 \subseteq \mathrm{VR}_2 \subseteq \mathrm{VR}_n$, the persistent or long-lived features are detected. Persistent features are analyzed through various topological summaries like Betti numbers, persistent diagrams, and persistent barcodes. For each filtration, the Betti numbers are identified such that Betti-0 (β_0) describes the number of connected components, Betti-1 (β_1) reports the number of one-dimensional holes or topological loops, and Betti-2 (β_2) gives the number of two-dimensional holes or topological voids, and so on. For a given filtration, persistent diagrams represent the feature with a point in the Cartesian coordinate system, with 'x' coordinate describing its birth time and 'y' coordinate describing its death time. The longer distance from main diagonal in persistent diagrams is treated as the notion of stronger persistence. Persistent barcodes capture the birth and death instances of the features as a bar.

3. PERSISTENT HOMOLOGY-BASED RESILIENCE ENHANCEMENT (PHRE) TECHNIQUE

A weighted graph $G = (V, E, w)$ describes the topology of the power grid with its buses comprising the set of vertices $V = (\nu_1, \nu_2, \cdots, \nu_n)$, its lines forming the set of edges $E = (\mathrm{e}_1, \mathrm{e}_2, \cdots, \mathrm{e}_m)$ and the edge weights w proportional to line impedances, or, admittances, or, power flows. The adjacent node connectivity is given by weighted adjacency matrix A_{pq} for the nodes p and q as

$$\mathrm{A}_{pq} = \begin{cases} w_{pq} & \forall \mathrm{e}_{pq} \in E, p \neq q \\ 0 & otherwise \end{cases} \tag{2}$$

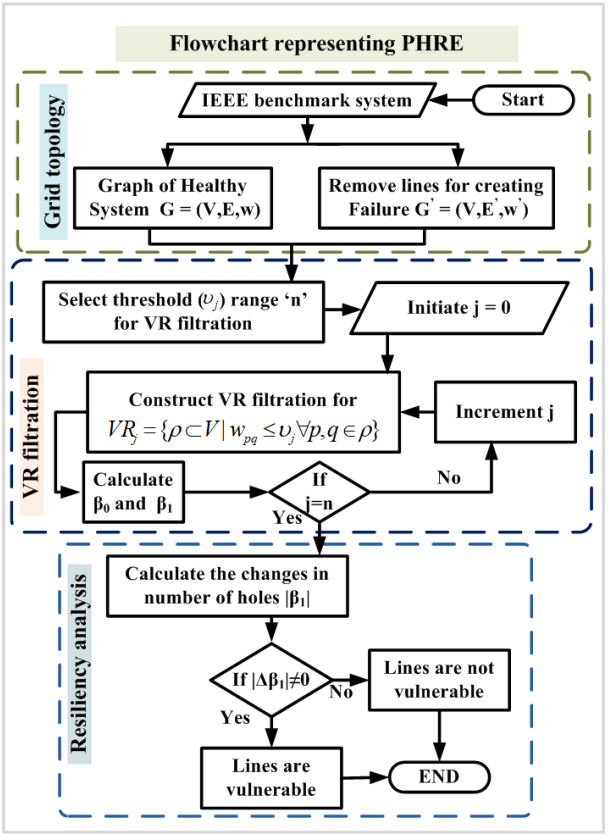


Fig. 1. The flowchart of proposed PHRE technique

The PHRE technique presented in Fig. 1 is applied for investigating the grid resiliency. The network filtration is carried out through VR filtration as explained in 1 for a range of thresholds $\upsilon_j, j = 0, \cdots, n$ imposed on edge weights. To obtain the Betti-0 (β_0) numbers, number of connected components are identified, and the presence of one-dimensional holes are detected to report the Betti-1 (β_1) numbers in the corresponding VR complex.

The blackouts due to cascading failures triggered by events like voltage instability, or overloading results into the change in topology as $G' = (V, E', w)$ whereas E' and w' are reduced set of edges and reduced vector of weights. The VR complices and corresponding betti numbers are identified for the similar range of weight thresholds considered previously. The change in betti numbers $\Delta\beta_0^{\upsilon_j}$ and $\Delta\beta_1^{\upsilon_j}$ for threshold υ_j is calculated as

$$|\Delta\beta_r^{\upsilon_j}| = \left|\beta_r^{b,\upsilon_j} - \beta_r^{f,\upsilon_j}\right| \tag{3}$$

where, $r = (0, 1)$, β_r^b corresponds to the healthy system and β_r^f corresponds to the deformed graph G' for failure.

The change in the number of one-dimensional holes $\left|\Delta\beta_1^{\upsilon_j}\right|$is used as the measure of the grid resiliency. The line outage changes the network topology and the deformation is reflected into the change in Betti numbers. The loss of edges will break the topological loops which results into fall in the number of holes i.e. β_1. If there is fall in the number of holes then the lines are more vulnerable as the outage has caused deformation to the network topology. If the number of holes are remain unchanged then it implies that the lines are not vulnerable.

4. REPRESENTATIVE CASE STUDY

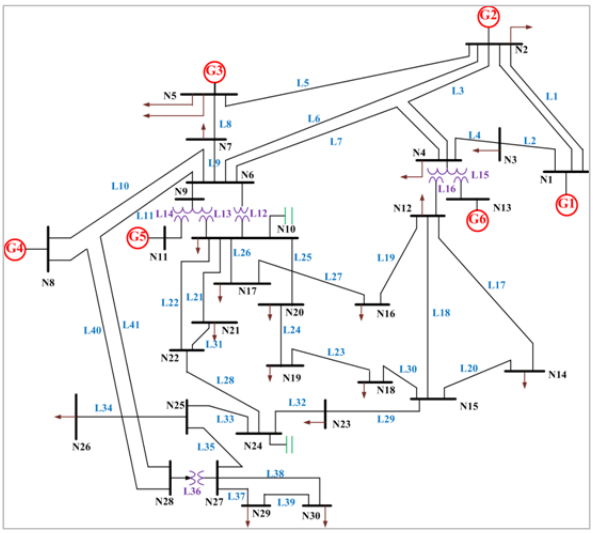


Fig. 2. One line diagram of benchmark IEEE 30 bus system

Benchmark IEEE 30 bus system (Fig.2) with 30 buses and 42 transmission lines of 289.1 MW generation and 283.4 MW load flow capacity Gupta et al. (2015) is considered as representative case study. Topological graph $G = (V, E, w)$ of this network is developed by considering the load and generator buses as the vertices, the transmission lines and transformers as the edges, and the impedance of the corresponding transmission line as edge weight. Two cases comprising base case corresponding to a healthy power grid and failure case corresponding to a cascade failure are discussed to verify the PHRE method.

4.1 Case 1: Base Case

Fig. 3 (h) shows the graph for base.The network filtration is computed with the VR filtration by varying the thresholds from $\upsilon = 0$ to $\upsilon = 0.7$ which are presented in Fig. 3 along with the Betti numbers associated with each complex. Fig. (a) describes the filtration with the threshold $\upsilon = 0$, where all the edges are discarded, and the point cloud of 30 nodes is observed. So the Betti-0 (β_0) = 30 and there are number holes so Betti-1 (β_1) = 0. With the increasing thresholds, more edges are getting included in the complices, and the number of β_0 is decreasing as more and more features are disappearing in the filtration. The one-dimensional holes are computed by tracing the topological loops present in the topology. The first hole appears at the threshold limit υ=0.2. With the further increase in the threshold, the number of holes, i.e., β_1, keeps on increasing. Finally, all edges are included at the threshold $\upsilon = 0.7$, and the last complex illustrates the healthy power grid.

4.2 Case 2: Failure Case

The cascading effect is created by tripping one of the lines from the base case system. The study has been conducted on the cascading scenario discussed in the Gupta et al. (2015) where the cascade has been initiated by tripping the line L21. Then the lines L22 and L29 tripped sequentially, resulting in cascade failure. The cascade scenario is simulated by removing the edges representing the lines L21, L22, and L29 from the graph of the base case. The VR complices are computed in a similar manner as the base case with the same threshold range. The changes in the Betti numbers for the failure case are demonstrated in Table 1. The number of connected components reported by the β_0 decrease with the varying threshold. It is observed that the first hole appears at the threshold $\upsilon = 0.2$, and the number of holes increases with further filtration.

Table 1. Betti numbers associated with VR complices for the cascade failure

Sr. No.	Threshold (υ)	Betti Numbers	
		β_0	β_1
1.	0.0	30	0
2.	0.2	11	1
3.	0.4	4	4
4.	0.6	2	4
5.	0.7	1	4

4.3 Comparative analysis

Table 2. Comparison of Betti numbers associated base case and failure case

Sr. No.	υ	Base Case		Failure Case		$\lvert\Delta\beta_0\rvert$	$\lvert\Delta\beta_1\rvert$
		β_0	β_1	β_0	β_1		
1.	0	30	0	30	0	0	0
2.	0.2	8	1	11	1	3	0
3.	0.4	3	5	4	4	1	1
4.	0.6	1	5	2	4	1	1
5.	0.7	1	6	1	4	0	2

Table 2 summarises the Betti numbers identified for the base case and failure case with the corresponding threshold limits. The variation in connected components (β_0) with the varying threshold for the base case and the failure case is illustrated in Fig. 4. The dynamics of one-dimensional holes (β_1) for various VR complices for the base case and the failure case is demonstrated in Fig. 5. It is observed that for the same threshold value, the number of one-dimensional holes β_1 drop down due to the cascade failure, and the number of β_0 increases as there are deformations in the network topology. Hence it is inferred that the number of holes is more for the healthy system, and during cascade failure, the number of holes drops down due to deformation of the topology. The more drastic change in the one-dimensional holes points out the fact that the corresponding lines are more vulnerable to failures. On the other hand, there may be line outages where the number of holes remains unchanged; such lines are not vulnerable to failure. Hence with the PHRE technique, it is deduced that the number of one-dimensional holes can be used as the measure of the resilience of the system.

5. CONCLUSION

The underlying geometry of power grid plays a vital role in investing its various functional characteristics for the resilience analysis. To gain profound insights into the vulnerability to the grid resilience, the TDA approach, specifically the persistent homology (PH) was explored. The extended abstract proposed an idea of a PHRE technique for analysing the capability of the system to retain its original

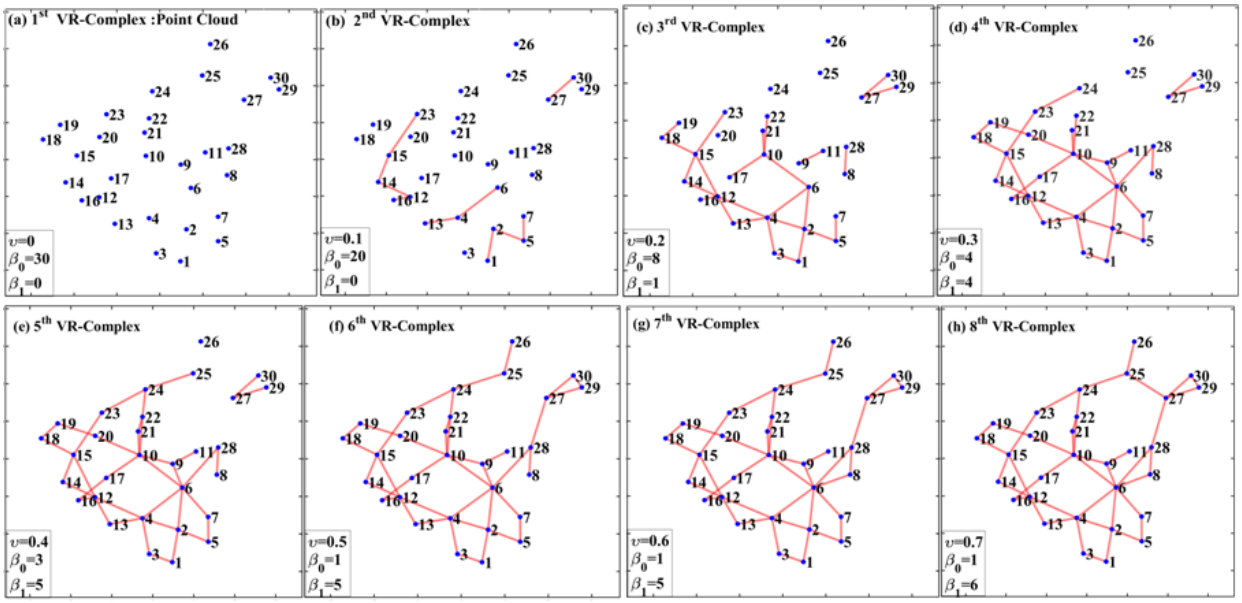


Fig. 3. VR complices of the base case with the threshold varying from $\upsilon = 0$ to $\upsilon = 0.7$ (The blue dots represent the vertices of the graph and the red lines represent the edges of the graph.)

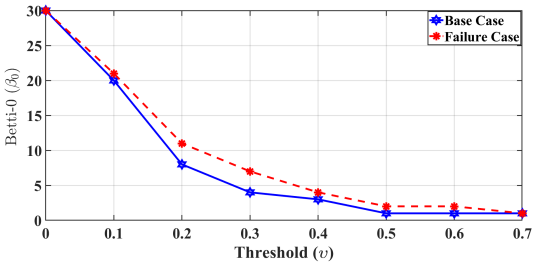


Fig. 4. Dynamics of Betti-0 (β_0) with respect to thresholds for base case and failure case

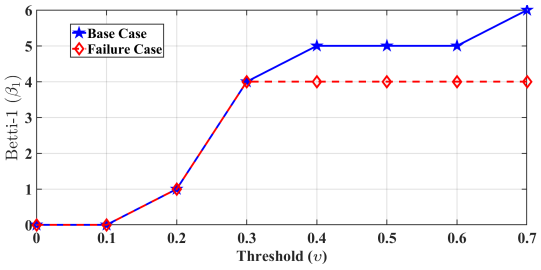


Fig. 5. Dynamics of Betti-1 (β_1) with respect to thresholds for base case and failure case

properties over a longer period under disruptive scenarios. The PHRE method was validated on the benchmark IEEE 30 bus system. From the results, it was observed that the healthy system reported a higher number of holes, and the cascade failure resulted in a drastic drop in the number of holes. Hence it was concluded that the number of holes or the topological loops might be treated as the measure of the system's ability to retain the resilience standard. The PHRE technique proved advantageous for analysis of cascade failure occurred by various factors, including component failure, overloading, voltage instability due to its relevance with the network geometry for characterizing its functional information.

REFERENCES

Cuadra, L., Salcedo-Sanz, S., Del Ser, J., Jiménez-Fernández, S., and Geem, Z.W. (2015). A critical review of robustness in power grids using complex networks concepts. *Energies*, 8(9), 9211–9265.

Dey, A.K., Gel, Y.R., and Poor, H.V. (2017). Motif-based analysis of power grid robustness under attacks. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 1015–1019. IEEE.

Gupta, S., Kambli, R., Wagh, S., and Kazi, F. (2015). Support-vector-machine-based proactive cascade prediction in smart grid using probabilistic framework. *IEEE Transactions on Industrial Electronics*, 62(4), 2478–2486.

Otter, N., Porter, M.A., Tillmann, U., Grindrod, P., and Harrington, H.A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6, 1–38.

Patania, A., Vaccarino, F., and Petri, G. (2017). Topological analysis of data. *EPJ Data Science*, 6(1), 1–6.

Sánchez-García, R.J., Fennelly, M., Norris, S., Wright, N., Niblo, G., Brodzki, J., and Bialek, J.W. (2014). Hierarchical spectral clustering of power grids. *IEEE Transactions on Power Systems*, 29(5), 2229–2237.

Schultz, P., Heitzig, J., and Kurths, J. (2014). Detours around basin stability in power networks. *New Journal of Physics*, 16(12), 125001.

Tajer, A., Perlaza, S.M., and Poor, H.V. (2021). *Advanced Data Analytics for Power Systems*. Cambridge University Press.

Zhu, Y., Yan, J., Tang, Y., Sun, Y.L., and He, H. (2014). Resilience analysis of power grids under the sequential attack. *IEEE Transactions on Information Forensics and Security*, 9(12), 2340–2354.

Zomorodian, A. (2010). Fast construction of the vietoris-rips complex. *Computers & Graphics*, 34(3), 263–271.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Hypergraph Based Distributed Quadratic Optimization

Ioannis Papastaikoudis * **Mengmou Li** * **Ioannis Lestas** *

* *Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ United Kingdom-emails:{ip352,ml995,icl20}@cam.ac.uk*

Abstract: We study a dual decomposition algorithm for a distributed optimization problem with a communication structure corresponding to a hypergraph. We prove that in the case of quadratic objective functions the respective discrete-time dynamical system of a modified dual decomposition algorithm that makes use of the Hessians of the objective functions converges in only one iteration.

Keywords: Large Scale Systems, Optimization: Theory and Algorithms.

1. INTRODUCTION

Real world networks usually consist of a large number of interconnected systems (agents) which communicate with each other in order to achieve a global objective. Many of these objectives can be formulated into distributed optimization problems. Distributed optimization can be traced back to the seminal works of [Tsitsiklis (1984)] and [Bertsekas and Tsitsiklis (1989)]. The most common way of solving distributed optimization problems is to use dual decomposition and first order methods (or subgradients) as in [Kelly et al. (1998)] and [Boyd et al. (2011)].

In this paper we study the discrete-time version of the dual decomposition case of a distributed optimization problem as it was presented in [Samar et al. (2007)]. This algorithm uses a hypergraph communication despite the fact that graph communication is the most commonly used in distributed optimization problems [Yang et al. (2019)]. Hypergraphs were introduced in [Berge (1973)] as a generalization of graphs. The importance of the hypergraph lies in the fact that it allows more than two nodes to be linked in the same edge (hyperedge). As a result, a hypergraph can depict more complex relationships compared to the communication structure of a graph. This different communication structure exists in reality e.g. large online social networks, supply chain management, etc. Various advantages of the hypergraph communication are presented in [Wolf et al. (2016)].

Our main contribution is to show that when the objective functions of the hypergraph distributed optimization problem are quadratic one can construct a dual decomposition based algorithm that converges in only one step, when the Hessians of the objective functions are available. This is in contrast to more classical first order methods where the convergence is asymptotic, i.e., this will involve an infinite number of steps.

An interesting observation is that the communication matrix of the dual decomposition algorithm is the Bolla's Laplacian for hypergraphs [Bolla (1993)]. This observation creates a link with graph based algorithms since they are most commonly described with the use of the graph Laplacian [Yang et al. (2019)]. In our setting the Laplacian communication matrix is also a projection matrix. As a result, the gap between the largest and smallest non zero eigenvalues is zero since both are equal to one. We know from [Duchi et al. (2011)] that a small aforementioned gap can lead to an improved convergence rate. This therefore hints that when the underlying communication topology is designed such that it corresponds to a hypergraph, optimization algorithms with faster convergence rate can be constructed.

The paper is organized as follows: Section 2 presents the mathematical tools that will be used. Section 3 introduces the distributed optimization problem in the hypergraph setting. Finally, in section 4 we present the dual decomposition algorithm along with the proposed modified version where the main results of the paper are associated with the convergence and the optimality of the modified algorithm.

2. PRELIMINARIES

2.1 Notation

The set of real numbers is $\mathbb{R}$. For $x \in \mathbb{R}^n, x \geq 0$ (resp. $x > 0$) means that all components of x are nonnegative (resp. positive). We define the cardinality of a set C as $|C|$. With $\text{vec}(a_i)_{i=1}^n$ we denote the vector obtained by stacking the vectors $a_1, ..., a_n$ into a column vector.

2.2 Convex Analysis

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if $f(x) - f(y) \leq \nabla f(x)^T(x - y)\ \forall\ x, y \in \mathbb{R}^n$. If strict inequality holds then f is a strictly convex function. A function $f : A \subset \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz if for each $x, x_0 \in A$, there exist constant $M > 0$ and $\delta_0 > 0$ such that $||x - x_0|| < \delta_0 \Rightarrow ||f(x) - f(x_0)|| \leq M||x - x_0||$.

2.3 Hypergraphs

A hypergraph (e.g. [Vitaly (2009)]) is a pair $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{v_1, ..., v_n\}$ is a finite set of nodes and $\mathcal{E} =$

$\{\mathcal{E}_1, ..., \mathcal{E}_m\}$ is the set of hyperedges. A hyperedge can join any number of nodes and not just two as it is in the case of a graph. The degree of a node in a hypergraph is the total number of hyperedges that are adjacent to this node. The size of a hyperedge $\mathcal{E}_j$, denoted by $|\mathcal{E}_j|$ is the total number of nodes that are adjacent to this hyperedge. We have that $|\mathcal{E}_j| \geq 2, \ \forall \ 1 \leq j \leq |\mathcal{E}|$. We define by D_V the diagonal $|\mathcal{V}| \times |\mathcal{V}|$ matrix whose entries are the degrees of each node and by D_E the diagonal $|\mathcal{E}| \times |\mathcal{E}|$ matrix whose diagonal entries are the sizes of each hyperedge. For a hypergraph $\mathcal{H}$, the incidence matrix denoted by E, is a $|\mathcal{V}| \times |\mathcal{E}|$ matrix whose (i, j)-th entry is given by

$$E_{ij} = \begin{cases} 1, & v_i \in \mathcal{E}_j \\ 0, & \text{otherwise.} \end{cases}$$

The Bolla's Laplacian for hypergraphs denoted by L, is given by

$$L = D_V - E(D_E)^{-1}E^T.$$

2.4 Matrix Theory

For a square matrix $M \in \mathbb{R}^{n\times n}$, $S(M)$ denotes the spectrum of matrix M. A square matrix $P \in \mathbb{R}^{n\times n}$ is called a projection matrix if $P^2 = P$. A square matrix $P \in \mathbb{R}^{n\times n}$ is called an orthogonal projection matrix if $P^2 = P = P^T$. For a projection matrix P its spectrum is $S(P) = \{0, 1\}$. A symmetric matrix M is called positive semidefinite $M \succeq 0$ if $x^T Mx \geq 0$ for every nonzero $x \in \mathbb{R}^n$. If we have strict inequality we say that matrix M is positive definite $M \succ 0$.

3. THE HYPERGRAPH DISTRIBUTED OPTIMIZATION ALGORITHM

We have the following distributed optimization problem with K subsystems

$$\min_{x=[x_1,...,x_K]} \quad \sum_{i=1}^{K} f_i(x_i)$$
$$\text{s.t.} \quad x_i = E_i z \quad i = 1, ..., K \tag{1}$$

where $f_i : \mathbb{R}^{p_i} \to \mathbb{R}$ is the objective function of ith subsystem and is considered to be strictly convex, continuously differentiable with its gradient ∇f_i being locally Lipschitz. The vectors $x_i, \forall \ 1 \leq i \leq K$ denote the variables of the K subsystems. For each subsystem i we have $x_i \in \mathbb{R}^{p_i}$ and we denote the lth component of vector x_i as x_i^l. We assume that all the components of all the variables appear in the variables of other subsystems [1] (i.e. they are coupling variables). We assume there are N different groups of coupling variables, with the variables in each group required to be equal. The components of vector $z \in \mathbb{R}^N$ give the respective common values in each of these groups. The relationship $x_i = E_i z$ defines the equality constraints ("consistency constraints") among the coupling variables of subsystem i and their respective common values. In particular, E_i is a $p_i \times N$ matrix whose (l, j)-th entry is given by

$$E_i^{lj} = \begin{cases} 1, & \text{if } \ x_i^l = z_j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Hence x_i^j is the coupling variable of the ith subsystem that belongs to the jth group of coupling variables for $i = 1, ..., K$ and $j = 1, ..., N$. The vector of coupling variables is denoted as $x = (x_1, ..., x_K) \in \mathbb{R}^p$, where $p = p_1 + ... + p_K$.

We consider a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ to represent the couplings among the variables. In particular, the set of nodes $\mathcal{V}$ is partitioned into $\mathcal{V} = \{\mathcal{V}_1, ..., \mathcal{V}_K\}$ where each node in subset $\mathcal{V}_i$ is associated with a component of variable x_i. Each hyperedge $\mathcal{E}_j$ is associated with a component z_j of vector z. Therefore the hyperedge set $\mathcal{E}$ is associated with the "consistency constraints" among the coupling variables of different subsystems. For the hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ we have $|\mathcal{V}| = p, |\mathcal{E}| = N$ and incidence matrix $E \in \mathbb{R}^{p\times N}$ where $E = \begin{bmatrix} E_1 \\ \vdots \\ E_K \end{bmatrix}$. The node degree matrix of the hypergraph is $D_V = I_{p\times p}$ and the hyperedge size matrix is $D_E = \text{diag}\{|\mathcal{E}_1|, ..., |\mathcal{E}_N|\}$ where $|\mathcal{E}_j| \geq 2, \forall \ 1 \leq j \leq N$. The relationship $x_i = E_i z, \forall \ 1 \leq i \leq K$ can also be written as $x = Ez$.

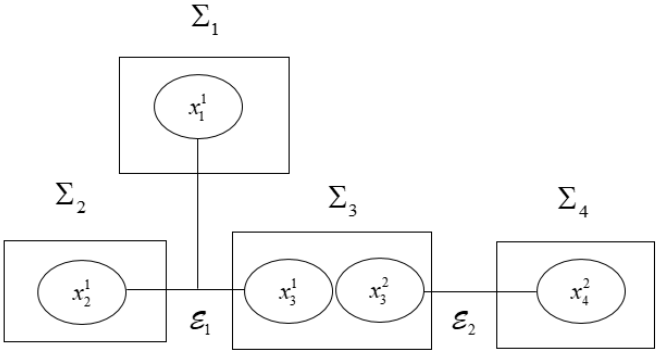


Fig. 1. Hypergraph Communication

It is important to note that this distributed optimization setting can be viewed as a multiple consensus problem, a consensus value must achieved for each hyperedge. The Lagrangian of (1) is

$$\mathcal{L}(x, z, v) = \sum_{i=1}^{K} f_i(x_i) - v^T(x - Ez) \tag{3}$$

where $v \in \mathbb{R}^p$ is the Lagrange multiplier associated with $x = Ez$. The equation (3) can also be written in the form

$$\mathcal{L}(x, z, v) = \sum_{i=1}^{K} (f_i(x_i) - v_i^T x_i) + v^T Ez$$

where v_i is the subvector of v associated with the ith subsystem. The optimality conditions are:

[1] Note that this is without loss of generality as minimization w.r.t. local variables that do not appear in other subsystems can be incorporated within the functions ∇f_i.

$$\frac{\partial L}{\partial x} = 0 \Rightarrow \nabla f_i = v_i \text{ (subsystems interconnections)} \quad (4a)$$
$$\frac{\partial L}{\partial v} = 0 \Rightarrow x = Ez \text{ (primal feasibility)} \quad (4b)$$
$$\frac{\partial L}{\partial z} = 0 \Rightarrow E^T v = 0 \text{ (dual feasibility).} \quad (4c)$$

Example 1. In Figure 1 we have a hypergraph communication with node set $\mathcal{V} = \{1, 2, 3, 4, 5\}$ and hyperedge set $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2\}$. For each node i there is a variable component x_i^j associated with it. Variables $\{x_1^1, x_2^1, x_3^1\}$ are attached to hyperedge $\mathcal{E}_1$ and variables $\{x_3^2, x_4^2\}$ are attached to hyperedge $\mathcal{E}_2$. The node degree matrix, hyperedge size matrix and the incidence matrix are

$$D_V = I_{5\times 5}, D_E = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } E = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \end{bmatrix} = \begin{pmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 \end{bmatrix} \end{pmatrix}$$

respectively.

Remark 1. It is important to note that in our setting $E^T E = D_E$.

4. DUAL DECOMPOSITION

In this section we will introduce the discrete-time dual decomposition algorithm and present also a modified version that leads to the main results in the paper.

To find the dual function we first minimize (3) over z, which results in the condition (4c). Then we solve the following subproblems independently for each i given v_i

$$\min_{x_i} f_i(x_i) - v_i^T x_i. \quad (5)$$

Since the objective functions $f_i(x_i)$ are strictly convex, the solution to each subsystem is unique. The dual of the original problem (1) is

$$\max_{v=[v_1,\ldots,v_K]} \sum_{i=1}^{K} q_i(v_i) \quad \text{s.t.} \quad E^T v = 0 \quad (6)$$

where $q_i(v_i) = \min_{x_i}\{f_i(x_i) - v_i^T x_i\}$. A detailed procedure for the extraction of the dual decomposition algorithm is given in [Samar et al. (2007)]. The algorithm is presented below:

Dynamical System

$$E^T v(0) = 0 \quad \text{initial conditions} \quad (7a)$$
$$\nabla f_i(x_i(k)) = v_i(k) \quad \text{optimize subsystems} \quad (7b)$$
$$z(k) = (E^T E)^{-1} E^T x(k) \quad \text{average } x \quad (7c)$$
$$v(k+1) - v(k) = -\rho(x(k) - Ez(k)) \quad \text{update } v \quad (7d)$$

where $k = 0, 1, 2, \ldots$ denotes the iteration number and $\rho > 0$ is the stepsize. Relationship (7b) results from (5). If we substitute (7c) in (7d) we have

$$v(k+1) - v(k) = -\rho(x(k) - E(E^T E)^{-1} E^T x(k)) \Rightarrow$$
$$v(k+1) - v(k) = -\rho Q x(k)$$

where

$$Q = I - E(E^T E)^{-1} E^T. \quad (9)$$

Remark 2. As mentioned in Remark 1, $E^T E = D_E$ and as a result, matrix Q in (9) can be written as $Q = D_V - E(D_E)^{-1} E^T$ since $D_V = I$. Hence matrix Q is the Bolla's Laplacian for the hypergraph considered, [Bolla (1993)]. It is proven in [Samar et al. (2007)] that matrix Q is an orthogonal projection matrix.

Quadratic Case Below we will present the main result of our work. We consider the problem (1) with the objective functions to be $f_i(x_i) = \frac{1}{2} x_i^T A_i x_i + b_i^T x_i + c_i$, $1 \le i \le K$. This problem can also be written in the following form,

$$\min_{x=[x_1,\ldots,x_K]} \frac{1}{2} x^T A x + b^T x + c \quad \text{s.t.} \quad x = Ez \quad (10)$$

where matrix $A = \text{diag}\{A_1, \ldots, A_K\}$ is a block diagonal matrix, each $A_i \in \mathbb{R}^{p_i \times p_i}, 1 \le i \le K$ is a symmetric positive definite matrix. Parameters b and c can be expressed as $b = \text{vec}(b_i)_{i=1}^K$ and $c = \sum_{i=1}^{K} c_i$ respectively. Below we propose a modification of dynamical system (7a)-(7d) which solves the optimization problem (10) in a single iteration.

Dynamical System

$$E^T v(0) = 0 \quad (11a)$$
$$\nabla f_i(x_i(k)) = v_i(k) \quad (11b)$$
$$z(k) = (E^T A E)^{-1} E^T A x(k) \quad (11c)$$
$$v(k+1) - v(k) = -A(x(k) - Ez(k)) \quad (11d)$$

Remark 3. We notice that a main difference of the dynamical system (11a)-(11d) compared to (7a)-(7d) is equation (11c). In equation (7c) we only average the primal variables while in equation (11c) we conduct a weighted averaging, where the weights are determined by the Hessians of the objective functions.

Remark 4. The choice $\rho = A$ in (11d) can be seen as using a non-uniform stepsize among subsystems.

Substituting (11c) in (11d) we have

$$v(k+1) - v(k) = -A(x(k) - E(E^T A E)^{-1} E^T A x(k)) \Rightarrow$$
$$v(k+1) - v(k) = -AQ' x(k)$$

where

$$Q' = I - E(E^T A E)^{-1} E^T A. \quad (13)$$

Lemma 4.1. Matrix Q' in (13) is a projection matrix.

Proof.

We have that

$$\begin{aligned}(Q')^2 =& [I - E(E^T A E)^{-1} E^T A]^2 \\ =& [I - E(E^T A E)^{-1} E^T A][I - E(E^T A E)^{-1} E^T A] \\ =& I - 2E(E^T A E)^{-1} E^T A \\ &+ E(E^T A E)^{-1} E^T A E (E^T A E)^{-1} E^T A \\ =& I - 2E(E^T A E)^{-1} E^T A + E(E^T A E)^{-1} E^T A \\ =& I - E(E^T A E)^{-1} E^T A \\ =& Q'.\end{aligned}$$

As a result, Q' is a projection matrix. The spectrum of Q' is $S(Q') = \{0, 1\}$.

Remark 5. It is important to note that matrix Q' is not symmetric and as a result, not an orthogonal projection matrix.

Using the fact that $\nabla f(x) = Ax+b = v \Rightarrow x = A^{-1}(v-b)$ the algorithm (11a)-(11d) can also be written compactly as follows:

$$E^T v(0) = 0 \tag{15a}$$

$$v(k+1) - v(k) = -AQ'A^{-1}v(k) + AQ'A^{-1}b. \tag{15b}$$

The following theorems present our main results which are associated with the convergence of the modified algorithm in one iteration and the optimality of its equilibrium point.

Theorem 4.2. The algorithm in (11a)-(11d) converges in one iteration.

Proof. We consider the compact form in (15a)-(15b). By considering an iteration to find $v(k+2)$ we have

$$\begin{aligned} v(k+2) =& (I - AQ'A^{-1})v(k+1) + AQ'A^{-1}b \\ =& (I - AQ'A^{-1})((I - AQ'A^{-1})v(k) + AQ'A^{-1}b) \\ &+ AQ'A^{-1}b \\ =& (I - AQ'A^{-1})^2 v(k) + (I - AQ'A^{-1})AQ'A^{-1}b \\ &+ AQ'A^{-1}b \\ =& (I - AQ'A^{-1})^2 v(k) + 2AQ'A^{-1}b \\ &- (AQ'A^{-1})^2 b. \end{aligned}$$

We notice that

$$\begin{aligned} (I - AQ'A^{-1})^2 =& (I - AQ'A^{-1})(I - AQ'A^{-1}) \\ =& I - 2AQ'A^{-1} + AQ'A^{-1}AQ'A^{-1} \\ =& I - 2AQ'A^{-1} + A(Q')^2A^{-1} \\ =& I - AQ'A^{-1} \end{aligned}$$

since Q' is a projection matrix which implies $(Q')^2 = Q'$. Similarly for $(AQ'A^{-1})^2$ we have that

$$(AQ'A^{-1})^2 = AQ'A^{-1}AQ'A^{-1} = A(Q')^2A^{-1} = AQ'A^{-1}.$$

As a result,

$$v(k+2) = (I - AQ'A^{-1})v(k) + AQ'A^{-1}b = v(k+1).$$

Therefore by induction $v(k) = v(1)$ for $k \geq 1$, hence the algorithm converges in one iteration.

Theorem 4.3. The algorithm in (11a)-(11d) satisfies the optimality conditions (4a)-(4c) at equilibrium.

Proof. Conditions (4a), (4b) are trivially satisfied at equilibrium from (11b), (11d), respectively. It remains to show that $E^T v = 0$ at equilibrium. Since we have the initial conditions $E^T v(0) = 0$ it is sufficient to show $E^T v(k+1) = E^T v(k)$ for all $k \geq 0$. Premultiplying (11d) by E^T we need to show that

$$E^T A(x(k) - Ez(k)) = 0.$$

This holds since

$$\begin{aligned} E^T A(x(k) - Ez(k)) &= E^T A(I - E(E^T AE)^{-1}E^T A)x(k) \\ &= 0. \end{aligned}$$

Remark 6. It should be noted that if A is diagonal then the iteration in Algorithm (11a)-(11d) decomposes into distributed computations within each hyperedge. If A is block diagonal distributed algorithms that terminate in a finite number of steps can be formulated by communicating the Hessians of the objective functions throughout the network. The development of more efficient distributed computational procedures in this case is part of ongoing work.

Remark 7. The classical first order algorithms usually require less information for their implementation (no Hessians, smaller connectivity) but they converge asymptotically, i.e., they tend to the optimal solution after an infinite number of steps.

5. CONCLUSION

We have considered a distributed optimization problem with coupling constraints formulated by means of a hypergraph. We have shown that when the objective functions are quadratic a modified dual decomposition algorithm, which makes use of the Hessians of the objective functions, can be constructed such that the optimal solution is obtained after one iteration. Future work includes extensions to problems with general convex objective functions where the hypergraph structure is exploited to obtain improved convergence rates.

REFERENCES

Berge, C. (1973). Graphs and hypergraphs.

Bertsekas, D. and Tsitsiklis, J. (1989). *Parallel and distributed computation: numerical methods*. Athena Scientific.

Bolla, M. (1993). Spectra, euclidean representations and clusterings of hypergraphs. *Discrete Mathematics*, 117(1-3), 19–39.

Boyd, S., Parikh, N., and Chu, E. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Duchi, J.C., Agarwal, A., and Wainwright, M.J. (2011). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3), 592–606.

Kelly, F.P., Maulloo, A.K., and Tan, D.K.H. (1998). Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3), 237–252.

Samar, S., Boyd, S., and Gorinevsky, D. (2007). Distributed estimation via dual decomposition. In *2007 European Control Conference (ECC)*, 1511–1516. IEEE.

Tsitsiklis, J. (1984). *Problems in decentralized decision making and computation*. Ph.D. thesis, Massachusetts Institute of Technology.

Vitaly, I. (2009). Voloshin. introduction to graph and hypergraph theory.

Wolf, M.M., Klinvex, A.M., and Dunlavy, D.M. (2016). Advantages to modeling relational data using hypergraphs versus graphs. In *2016 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–7. IEEE.

Yang, T., Yi, X., Wu, J., Yuan, Y., Wu, D., Meng, Z., Hong, Y., Wang, H., Lin, Z., and Johansson, K.H. (2019). A survey of distributed optimization. *Annual Reviews in Control*, 47, 278–305.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A note on explicit data-driven (M)PC

Manuel Klädtke * **Dieter Teichrib** * **Nils Schlüter** *
Moritz Schulze Darup *

* *Control and Cyberphysical Systems Group,*
TU Dortmund University, Germany
(e-mails: {manuel.klaedtke, moritz.schulzedarup}@tu-dortmund.de)

Abstract: We show that the explicit solution of a data-driven predictive control scheme for deterministic LTI systems may not be as intractable as previously assumed. By comparing the structure of resulting parametric quadratic programs for the data-driven and classical model-based formulation, we analyze similarities and redundancies that ultimately lead to related structures of the respective explicit solutions. More precisely, some observations indicate a one-to-one relationship of these solutions that will be explored in future work. We illustrate this result by a thorough analysis of a simple example.

Keywords: Data-driven control, model predictive control, explicit MPC, parametric optimization

1. INTRODUCTION

Data-driven predictive control (DPC), where the prediction of the systems' behavior is carried out based on collected input-output data instead of a model, is becoming more and more popular (see, e.g., Coulson et al. (2019); Berberich and Allgöwer (2020); Dörfler et al. (2021)). Remarkably, assuming perfect data and linear dynamics, the fundamental lemma by Willems et al. (2005) and variants of it (as proposed by van Waarde et al. (2020) and Markovsky and Dörfler (2020)) allow establishing the equivalence of the data-driven and model-based approach with respect to the resulting control actions.

However, while strongly related, the two approaches lead to different optimal control problems (OCP). In fact, DPC usually results in an OCP with significantly more decision variables than model-based predictive control (MPC, see Rawlings et al. (2017) for an overview). Moreover, MPC typically results in strictly convex OCP while (unmodified) DPC only offers convexity. As a consequence, explicit solutions of the data-driven OCP seem "unattractive" at first sight (Alpago et al., 2020, Section IV.B), even for applications where explicit MPC (Bemporad et al., 2002) is tractable.

We show in this note that the perceived imbalance between DPC and MPC can be resolved in some cases. In fact, while more decision variables indeed typically result in more complex explicit solutions (in terms of the number of regions etc.), this is not the case for deterministic DPC in comparison to its model-based analogue. To see this, we establish a novel and stricter relation between the two approaches, which reveals that the larger number of decision variables merely results in ambiguity rather than more complex explicit solutions. Note that previous approaches to explicit DPC either operate in the statespace, i.e. assume fully measurable states (Sassella et al., 2021), or utilize additional regularization to obtain a strictly convex OCP (Breschi et al., 2021). For our analysis, we will not apply such simplifications or modifications.

The note is organized as follows. In Section 2, we summarize classical MPC and fundamentals of DPC. The analysis of explicit solutions of the corresponding OCPs and the identification of a closer relation between them are carried out in Section 3. We illustrate our findings with an illustrative example in Section 4. Finally, promising directions for future research are discussed in Section 5.

2. FUNDAMENTALS OF MPC AND DPC

2.1 Classical MPC

We briefly summarize classical MPC in a form that is compatible with the data-driven realization in Section 2.2. To this end, we assume that a linear prediction model

$$x(k+1) = Ax(k) + Bu(k) \tag{1a}$$
$$y(k) = Cx(k) + Du(k) \tag{1b}$$

is known. We further assume that input and output constraints are given in terms of polyhedral sets

$$\mathcal{U} := \{u \in \mathbb{R}^m \,|M_u u \leq v_u\}, \mathcal{Y} \quad := \{y \in \mathbb{R}^p \;|M_y y \leq v_y\},$$

which are specified by the matrices $M_{u/y}$ and vectors $v_{u/y}$, respectively. Then, classical MPC (without terminal cost and constraints) can be realized by solving

$$\begin{aligned} \min_{u(k),x(k),y(k)} \; & \sum_{k=0}^{N_f-1} \|y(k)\|_Q^2 + \|u(k)\|_R^2 && (2)\\ \text{s.t.} \quad & x(0) = x_0, \\ & x(k+1) = A\,x(k) + Bu(k), && \forall k \in \{0,...,N_f-2\}, \\ & y(k) = C\,x(k) + Du(k), && \forall k \in \{0,...,N_f-1\}, \\ & (y(k),u(k)) \in \mathcal{Y} \times \mathcal{U}, && \forall k \in \{0,...,N_f-1\} \end{aligned}$$

in every time-step for the current state x_0, where Q and R are assumed to be positive definite weighting matrices and where $N_f \in \mathbb{N}$ denotes the prediction horizon. Now, the OCP (2) is typically condensed into a quadratic program (QP) such that only the inputs remain as decision variables. To this end, one first introduces the sequences

$$\mathrm{u}_f := \begin{pmatrix} u(0) \\ \vdots \\ u(N_f-1) \end{pmatrix} \quad \text{and} \quad \mathrm{y}_f := \begin{pmatrix} y(0) \\ \vdots \\ y(N_f-1) \end{pmatrix}, \tag{3}$$

and the extended weighting matrices $\mathcal{Q} := \operatorname{diag}(Q, \dots, Q)$ and $\mathcal{R} := \operatorname{diag}(R, \dots, R)$ in order to rewrite the cost function as

$$\sum_{k=0}^{N_f-1} \|y(k)\|_Q^2 + \|u(k)\|_R^2 = \|\mathrm{y}_f\|_{\mathcal{Q}}^2 + \|\mathrm{u}_f\|_{\mathcal{R}}^2. \tag{4}$$

We further define the matrices

$$\mathcal{O}_N := \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{N-1} \end{pmatrix} \text{ and } \mathcal{T}_N := \begin{pmatrix} D & & 0 \\ CB & \ddots & \\ \vdots & \ddots & \ddots \\ CA^{N-2}B & \dots & CB \;\; D \end{pmatrix},$$

which we will consider for different N during this note. For $N = N_f$, we then obtain the relation

$$\mathrm{y}_f = \mathcal{O}_{N_f} x_0 + \mathcal{T}_{N_f} \mathrm{u}_f. \tag{5}$$

Finally, substituting (5) into (4) and introducing the augmented matrices $\mathcal{M}_{u/y} := \operatorname{diag}\left(M_{u/y}, \dots, M_{u/y}\right)$ leads to

$$\begin{aligned} \mathrm{u}_f^*(x_0) := \arg\min_{\mathrm{u}_f} \; & \frac{1}{2} \mathrm{u}_f^\top H \mathrm{u}_f + x_0^\top F^\top \mathrm{u}_f \\ \text{s.t.} \quad & G\mathrm{u}_f \le E x_0 + d \end{aligned} \tag{6}$$

with the parameter x_0 as well as

$$\begin{aligned} H &:= 2\mathcal{T}_{N_f}^\top \mathcal{Q} \mathcal{T}_{N_f} + 2\mathcal{R}, & F &:= 2\mathcal{T}_{N_f}^\top \mathcal{Q} \mathcal{O}_{N_f}, \\ G &:= \begin{pmatrix} \mathcal{M}_u \\ \mathcal{M}_y \mathcal{T}_{N_f} \end{pmatrix} & E &:= \begin{pmatrix} 0 \\ -\mathcal{M}_y \mathcal{O}_{N_f} \end{pmatrix}, \\ d &:= \begin{pmatrix} v_u^\top & \dots & v_u^\top & v_y^\top & \dots & v_y^\top \end{pmatrix}^\top. \end{aligned} \tag{7}$$

2.2 DPC using input-output sequences

In contrast to MPC, DPC considers input-output data instead of a model as in (1). More precisely, DPC builds (in its simplest form) on two sequences u_d and y_d as in (3) but of length $N_d \in \mathbb{N}$ that reflect prerecorded system inputs and outputs. We note, at this point, that with slight abuse of notation, we denote both the elements of u_f and u_d with $u(k)$. However, the relationship will always be clear from the context. The same applies to the elements of the sequences y_f and y_d. Now, in order to realize DPC by means of u_d and y_d, the sequences have to carry enough information about the systems' dynamics. This condition can be fulfilled if the output sequence y_d is consistent with a persistently exciting and sufficiently long input sequence u_d. More specifically, for deterministic DPC, which we consider here, consistency means that there exists a model (1) with initial state x_0 such that

$$\mathrm{y}_d = \mathcal{O}_{N_d} x_0 + \mathcal{T}_{N_d} \mathrm{u}_d. \tag{8}$$

Further, according to Willems et al. (2005), u_d is persistently exciting of order $N_e \in \mathbb{N}$ if the Hankel matrix

$$\mathcal{H}_{N_e}(\mathrm{u}_d) := \begin{pmatrix} u(0) & u(1) & \dots & u(N_d - N_e) \\ u(1) & u(2) & \dots & u(N_d - N_e + 1) \\ \vdots & & \ddots & \vdots \\ u(N_e - 1) & u(N_e) & \dots & u(N_d - 1) \end{pmatrix}$$

has full row rank, i.e., $\operatorname{rank}(\mathcal{H}_{N_e}(\mathrm{u}_d)) = mN_e$. Note that this requires $\mathcal{H}_{N_e}(\mathrm{u}_d)$ to have as least as many columns as rows, i.e.,

$$N_d - N_e + 1 \ge mN_e \quad \Longleftrightarrow \quad N_d \ge (m+1)N_e - 1. \tag{9}$$

Finally, the fundamental lemma by Willems et al. (2005) allows associating the given sequences u_d and y_d with other input-output sequences of the same system. In fact, under the assumption that the underlying system is linear, controllable, and u_d is persistently exciting of order $N_e := N_c + n$, with n being the state dimension, candidate sequences $(\mathrm{u}_c, \mathrm{y}_c)$ of length $N_c \in \mathbb{N}$ belong to the same system as $(\mathrm{u}_d, \mathrm{y}_d)$ if and only if

$$\begin{pmatrix} \mathrm{u}_c \\ \mathrm{y}_c \end{pmatrix} \in \operatorname{im} \begin{pmatrix} \mathcal{H}_{N_c}(\mathrm{u}_d) \\ \mathcal{H}_{N_c}(\mathrm{y}_d) \end{pmatrix}.$$

At this point, we briefly note that recent extensions of the fundamental lemma by van Waarde et al. (2020) and Markovsky and Dörfler (2020) allow alleviating some of the restrictions above. Now, in order to utilize the previous results for DPC, we proceed similarly to Coulson et al. (2019). We choose an integer $N_p \le n$ equal to or larger than the observability index, i.e., such that the corresponding matrix $\mathcal{O}_{N_p}$ has full column rank, which clearly requires observability. Next, we assume that u_d is persistently exciting of order

$$N_e := N_p + N_f + n. \tag{10}$$

According to the fundamental lemma, we then find that the concatenated sequences $(\mathrm{u}_p^\top \;\; \mathrm{u}_f^\top)^\top$ and $(\mathrm{y}_p^\top \;\; \mathrm{y}_f^\top)^\top$ with

$$\mathrm{u}_p := \begin{pmatrix} u(-N_p) \\ \vdots \\ u(-1) \end{pmatrix} \quad \text{and} \quad \mathrm{y}_p := \begin{pmatrix} y(-N_p) \\ \vdots \\ y(-1) \end{pmatrix}$$

and with $(\mathrm{u}_f, \mathrm{y}_f)$ as in (3), belong to the same system as $(\mathrm{u}_d, \mathrm{y}_d)$ if and only if

$$\begin{pmatrix} \mathrm{u}_p \\ \mathrm{u}_f \\ \mathrm{y}_p \\ \mathrm{y}_f \end{pmatrix} = \begin{pmatrix} \mathcal{H}_{N_p+N_f}(\mathrm{u}_d) \\ \mathcal{H}_{N_p+N_f}(\mathrm{y}_d) \end{pmatrix} a. \tag{11}$$

for some $a \in \mathbb{R}^l$ with $l := N_d - N_f - N_p + 1$. Based on reordering and partitioning, (11) can be rewritten as

$$\xi := \begin{pmatrix} \mathrm{u}_p \\ \mathrm{y}_p \end{pmatrix} = W_p a, \quad \mathrm{u}_f = U_f a, \quad \text{and} \quad \mathrm{y}_f = Y_f a \tag{12}$$

with the matrices W_p, U_f, and Y_f representing blocks of the concatenated Hankel matrices. We are now ready to formulate the OCP associated with DPC. In fact, the combination of (4) and (12) allow expressing the costs

$$\|\mathrm{y}_f\|_{\mathcal{Q}}^2 + \|\mathrm{u}_f\|_{\mathcal{R}}^2 = \|a\|_{Y_f^\top \mathcal{Q} Y_f + U_f^\top \mathcal{R} U_f}^2$$

as a function of a. Taking into account the constraints $\mathcal{M}_u \mathrm{u}_f \le \mathcal{V}_u$ and $\mathcal{M}_y \mathrm{u}_f \le \mathcal{V}_y$ and the remaining condition $\xi = W_p a$ then leads to the QP

$$\begin{aligned} a^*(\xi) := \arg\min_{a} \; & \frac{1}{2} a^\top \tilde{H} a \\ \text{s.t.} \quad & \tilde{G} a \le d, \\ & W_p a = \xi \end{aligned} \tag{13}$$

with the parameter ξ, the vector d as in (6), and

$$\tilde{H} := 2Y_f^\top \mathcal{Q} Y_f + 2U_f^\top \mathcal{R} U_f, \qquad \tilde{G} := \begin{pmatrix} \mathcal{M}_u U_f \\ \mathcal{M}_y Y_f \end{pmatrix}.$$

Remarkably, the role of the initial state x_0 in (6) is replaced by ξ, i.e., the N_p previous inputs and outputs, in (13). Furthermore, $a^*(\xi)$ only reflects an intermediate result that is used to compute optimal inputs via $\mathrm{u}_f^*(\xi) := U_f a^*(\xi)$.

3. FROM EXPLICIT MPC TO EXPLICIT DPC

The QPs (6) or (13) are typically solved for the current state x_0 or the most recent sequences ξ, respectively, to

obtain the optimal input for the current time-step. Subsequently, the procedure is repeated at the next sampling instance. Alternatively, in order to avoid numerical optimization during runtime, (6) can also be solved explicitly using parametric optimization. As a result, we then find the continuous and piecewise affine (PWA) solution

$$\mathrm{u}_f^*(x_0) = \begin{cases} L_1 x_0 + c_1 \text{ if } & x_0 \in \mathcal{X}_1, \\ \vdots & \vdots \\ L_s x_0 + c_s \text{ if } & x_0 \in \mathcal{X}_s, \end{cases} \tag{14}$$

which is defined on a polyhedral partition $\{\mathcal{X}_i\}_{i=1}^s$ of the state space (Bemporad et al., 2002). Computing this solution offline and evaluating it online is referred to as explicit MPC. While conceptually attractive, explicit MPC can usually only be applied for moderate "sizes" of the underlying QP since it is well-known that the number of regions $s \in \mathbb{N}$ typically grows exponentially with the number of decisions variables and constraints. As a consequence, solving (13) parametrically seems unattractive at first sight, since especially the number of decision variables is significantly larger than in (6). In fact, while u_f is of dimension mN_f, the dimension l of a is lower-bounded by

$$\begin{aligned} l &\geq (m+1)(N_p + N_f + n) - N_f - N_p \\ &= mN_p + mN_f + (m+1)n \end{aligned}$$

according to (9) and (10). Now, while the difference of at least $mN_p + (m+1)n$ decisions variables is significant especially for $m > 1$, we claim that this increase does not result in a more complex explicit solution for the special case of deterministic DPC. In fact, we claim that the increase in decision variables only leads to ambiguous solutions and that this ambiguity can be reduced (or even removed) by systematically eliminating variables.

3.1 Eliminating equality constraints for DPC

Following this claim, we initially eliminate the equality constraints in (13). To this end, we assume that a generalized inverse W_p^+ of W_p (satisfying the Penrose conditions) and a matrix V_p characterizing the null-space of W_p (i.e., $\mathrm{im}(V_p) = \ker(W_p)$) are known. Then, we can substitute a in (13) with

$$a := W_p^+ \xi + V_p \alpha,$$

where α is of dimension $\nu := \mathrm{nullity}(W_p) = l - \mathrm{rank}(W_p)$. Clearly, the equality constraints in (13) are satisfied for every $\alpha \in \mathbb{R}^\nu$. Hence, we obtain the transformed QP

$$\begin{aligned} \alpha^*(\xi) = \arg\min_{\alpha} \ & \frac{1}{2}\alpha^\top \hat{H} \alpha + \xi^\top \hat{F}^\top \alpha \\ \text{s.t.} \quad & \hat{G}\alpha \leq \hat{E}\xi + d \end{aligned} \tag{15}$$

with

$$\hat{H} := V_p^\top \tilde{H} V_p,\ \hat{G} := \tilde{G} V_p,\ \hat{F} := V_p^\top \tilde{H} W_p^+,\ \hat{E} := -\tilde{G} W_p^+.$$

Now, taking into account that W_p contains mN_p rows of the full rank matrix $\mathcal{H}_{N_p+N_f}(\mathrm{u}_d)$, we obviously have $\mathrm{rank}(W_p) \geq mN_p$ (and tighter bounds can be obtained with moderate effort). Hence, the number of decision variables in (15) is significantly smaller than in (13). Apart from this benefit, the two OCPs are almost equivalent. However, the transition from (13) to (15) involves a subtle modification. In fact, while (13) is infeasible for ξ not belonging to the system, (15) may be feasible for such ξ. This is due to to the fact that $W_p W_p^+ \xi$ (implicitly) maps such ξ to belonging ones as specified in Corollary 1 below.

3.2 The "common denominator"

The QP (15) not only provides a reduced number of decision variables but it also allows establishing a closer relation between MPC and DPC. To see this, we simply need to specify the (trivial) relation between x_0 and ξ. To this end, we recognize that ξ and the assumed observability allow reconstructing $x(-N_p)$. Using u_p, it is then easy to derive x_0. More formally, the procedure is captured by the relation $x_0 = \Gamma \xi$, where

$$\Gamma := \begin{pmatrix} \mathcal{B} - A^{N_p} \mathcal{O}_{N_p}^+ \mathcal{T}_{N_p} & A^{N_p} \mathcal{O}_{N_p}^+ \end{pmatrix}$$

with $\mathcal{B} := \begin{pmatrix} A^{N_p-1}B & \dots & B \end{pmatrix}$ and $\mathcal{O}_{N_p}^+ := (\mathcal{O}_{N_p}^\top \mathcal{O}_{N_p})^{-1} \mathcal{O}_{N_p}^\top$. Using this relation, we obtain the following relationship between the QPs (6) and (15).

Corollary 1. The cost and constraint specifications of (6) and (15) satisfy

$$\begin{aligned} \hat{H} &= V_p^\top U_f^\top H U_f V_p, \\ \hat{F} &= V_p^\top U_f^\top F \Gamma W_p W_p^+ + V_p^\top U_f^\top H U_f W_p^+, \\ \hat{G} &= G U_f V_p, \quad \hat{E} = E\Gamma W_p W_p^+ - G U_f W_p^+. \end{aligned}$$

A formal proof is omitted due to space restrictions. We note, however, that the key ingredient is the equation

$$Y_f = \mathcal{O}_{N_f} \Gamma W_p + \mathcal{T}_{N_f} U_f, \tag{16}$$

which can be considered as a variant of (5) for multiple sequences represented by Hankel matrices and which similarly appears in Persis and Tesi (2020). In fact, given (16), the relations in Corollary 1 follow from simple rearrangements and facts like $W_p V_p = 0$.

By further exploiting the novel relation between (6), (13), and (15), one can derive even stricter connections between MPC and DPC. In fact, we show in Klädtke et al. (2022) that there is actually a one-to-one relation between the (explicit) solutions of (6) and (15). While a formal specification of this relation, which is based on an additional parametrization and analysis of the involved subspaces, is beyond the scope of this note, we illustrate some of our findings with a numerical example next.

4. CASE STUDY

For our case study, we consider system (1) with

$$A = 1.2 \quad \text{and} \quad B = C = D = 1.$$

and the constraints $\mathcal{U} = [-1, 1]$ and $\mathcal{Y} = [-4, 4]$. Further, we choose $Q = R = 0.5$ and $N_f = 2$, which already determines the MPC problems (2) and (6). Explicitly solving (6) then leads to (14), where the $s = 5$ segments are specified by

$$L_1 = L_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad L_2 = L_4 = \begin{pmatrix} 0 \\ -0.6 \end{pmatrix}, \quad L_3 = \begin{pmatrix} -0.64 \\ -0.28 \end{pmatrix},$$

$$c_1 = -c_5 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad c_2 = -c_4 = \begin{pmatrix} 1 \\ -0.5 \end{pmatrix}, \quad c_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

as well as $\mathcal{X}_1 = [-5, -2.5]$, $\mathcal{X}_2 = [-2.5, -1.5625]$, $\mathcal{X}_3 = [-1.5625, 1.5625]$, $\mathcal{X}_4 = [1.5625, 2.5]$, and $\mathcal{X}_5 = [2.5, 5]$. The corresponding PWA solution is illustrated in Figure 1.

Now, to setup and investigate the DPC, we first note that $N_p = 1$ guarantees full rank of $\mathcal{O}_{N_p} = C = 1$. Hence, we choose an input sequence u_d, which is persistently exciting

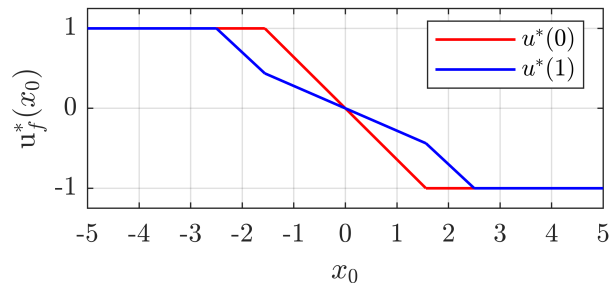


Fig. 1. Explicit solution $\mathrm{u}_f^*(x_0)$ for MPC.

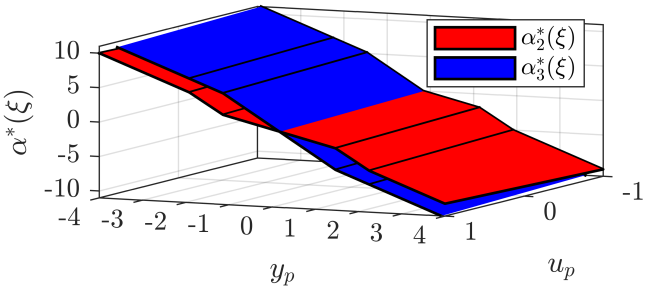


Fig. 2. Explicit solution $\alpha^*(\xi)$ for (modified) DPC. Note that (y_p, u_p) is artificially restricted to $\mathcal{Y} \times \mathcal{U}$ for visualization.

of order $N_e = 4$. According to (9), this requires at least $N_d = 7$ elements. It can be easily verified that

$$\mathrm{u}_d := (-0.6 \;\; 0 \;\; 0 \;\; 0 \;\; 0.5 \;\; 0.5 \;\; 1)^\top$$

satisfies all conditions. Furthermore,

$$\mathrm{y}_d := (-0.1 \;\; 0 \;\; 0 \;\; 0 \;\; 0.5 \;\; 1 \;\; 2.1)^\top$$

is a consistent output sequence since (8) is satisfied for $x_0 = 0.5$. According to (12), u_d and y_d specify

$$W_p = \begin{pmatrix} -0.6 & 0 & 0 & 0 & 0.5 \\ -0.1 & 0 & 0 & 0 & 0.5 \end{pmatrix}, \quad U_f = \begin{pmatrix} 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 1 \end{pmatrix},$$

and Y_f with $l = 5$. In the following, we mainly focus on the transformation to (15) and its explicit solution. To this end, we require the generalized inverse W_p^+ and the null-space description via V_p. Taking $\operatorname{rank}(W_p) = 2$ and, consequently, $\nu = 3$ into account, suitable choices are

$$W_p^+ = \begin{pmatrix} -2 & 2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ -0.4 & 2.4 \end{pmatrix} \quad \text{and} \quad V_p = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Obviously, the transformation from (13) to (15) allows to reduce the number of decision variables from $l = 5$ to $\nu = 3$. Still, (6) involves only $mN_f = 2$ variables. Hence, following our original claim, the transformed DPC problem should still offer some ambiguity. This can be confirmed by investigating the matrices

$$\hat{H} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.50 & 0.75 \\ 0 & 0.75 & 1.75 \end{pmatrix}, \quad \hat{G} = \begin{pmatrix} 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & -0.5 \\ 0 & -0.5 & -0.5 \\ 0 & 0 & 0.5 \\ 0 & 0.5 & 1 \\ 0 & 0 & -0.5 \\ 0 & -0.5 & -1 \end{pmatrix},$$

$$\hat{F} = \begin{pmatrix} 0 & 0 \\ -0.62 & 3.72 \\ -1.34 & 8.04 \end{pmatrix},$$

In fact, as apparent from the zero rows and columns, the choice of α_1 neither affects the cost function nor the constraint satisfaction. Without giving details, this is not surprising since $(1 \;\; 0 \;\; 0)^\top$ spans the null-space of $U_f V_p$. Now, eliminating this variable from (13) results in a strictly convex QP (whereas (13) is only convex as apparent from $\hat{H}$). Solving the reduced QP explicitly leads to the PWA functions in Figure 2. Clearly, while the domain is different, $\alpha^*(\xi)$ likewise consists of $s = 5$ affine segments.

5. CONCLUSIONS AND OUTLOOK

By establishing a stricter relation to its explicit MPC counterpart, we have shown that explicit DPC for deterministic LTI systems may not be as intractable as the parameter dimensions of the corresponding OCP suggest. While the analyzed example is kept simple to allow visualization, similar results can be obtained for more complex systems.

Future work will, among other things, thoroughly establish the connection between both formulations in terms of the (number of) partitions and (strongly) active constraints as well as further classify and reduce the amount of redundancies. Furthermore, it must be examined how (well) the results can be extended to systems or data with uncertainty.

REFERENCES

Alpago, D., Dörfler, F., and Lygeros, J. (2020). An extended Kalman filter for data-enabled predictive control. *IEEE Control Systems Letters*, 4, 994–999.

Bemporad, A., Morari, M., Dua, V., and Pistikopoulos, E.N. (2002). The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1), 3–20.

Berberich, J. and Allgöwer, F. (2020). A trajectory-based framework for data-driven system analysis and control. *2020 European Control Conference*, 1365–1370.

Breschi, V., Sassella, A., and Formentin, S. (2021). On the design of regularized explicit predictive controllers from input-output data. arXiv:2110.11808v1

Coulson, J., Lygeros, J., and Dörfler, F. (2019). Data-enabled predictive control: In the shallows of the DeePC. *2019 European Control Conference*, 307–312.

Dörfler, F., Coulson, J., and Markovsky, I. (2021). Bridging direct&indirect data-driven control formulations via regularizations and relaxations. arXiv:2101.01273v2.

Klädtke, M., Teichrib, D., Schlüter, N., and Schulze Darup, M. (2022). A deterministic view on explicit data-driven (M)PC. arXiv:2206.07025v2

Markovsky, I. and Dörfler, F. (2020). Identifiability in the behavioral setting. Available at `http://homepages.vub.ac.be/imarkovs/publications/identifiability.pdf`

Persis, C.D. and Tesi, P. (2020). Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, 65, 909–924.

Rawlings, J.B., Mayne, D.Q., and Diehl, M.M. (2017). *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2nd edition edition.

Sassella, A., Breschi, V., and Formentin, S. (2021) Learning explicit predictive controllers: theory and applications. arXiv:2108.08412v2

van Waarde, H.J., Persis, C.D., Çamlibel, M.K., and Tesi, P. (2020). Willems' fundamental lemma for state-space systems and its extension to multiple datasets. *IEEE Control Systems Letters*, 4, 602–607.

Willems, J.C., Rapisarda, P., Markovsky, I., and De Moor, B.L. (2005). A note on persistency of excitation. *Systems & Control Letters*, 54(4), 325–329.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Performance Estimation of First-Order Methods on Quadratic Functions ⋆

Nizar Bousselmi * **Julien M. Hendrickx** * **François Glineur** *

* *ICTEAM, UCLouvain, 1348 Louvain-la-Neuve, Belgium (e-mail: {nizar.bousselmi,julien.hendrickx,francois.glineur}@uclouvain.be).*

Abstract: We are interested in determining the worst performance exhibited by a given first-order optimization method on the class of quadratic functions. Since its introduction, the *Performance Estimation Problem* (PEP) methodology has allowed the computation of the exact worst-case performance of first-order optimization methods on several functions classes, including smooth convex, strongly convex or nonconvex functions.
In this work, we extend the PEP framework to the class of quadratic functions, and apply it to analyze the difference of performance of the gradient method between convex quadratic and general smooth convex functions.

Keywords: Performance estimation, First-order methods, Quadratic functions, Linear matrix inequalities

1. INTRODUCTION

The *Performance Estimation Problem* (PEP) methodology (introduced by Drori and Teboulle (2014)) allows to compute the exact worst-case performance of a first-order optimization method on a given class of functions. More precisely, given a method and a performance criterion (lower is better), a PEP is an optimization problem that maximizes this criterion among all possible functions belonging to some class. Thus, it provides the worst possible behavior of the method on the class of functions.

It has been shown in Taylor et al. (2017) that a PEP can be reformulated as a convex semidefinite program for a wide range of function classes $\mathcal{C}$. This provided several tight results on the performance of first-order methods. In particular, the worst-case behavior of the *Gradient Method* (GM) on the class $\mathcal{F}_{\mu,L}$ of L-smooth μ-strongly convex functions was exhaustively covered.

In this work, we extend the PEP framework to function classes defined by matrices. This typically allows to study the worst-case performance of first-order methods on the class $\mathcal{Q}_{\mu,L}$ of homogeneous quadratic functions of the form $f(x) = \frac{1}{2}x^TQx$ with $\mu I \preceq Q \preceq LI$ for given parameters μ and L ($0 \leq \mu \leq L$). Another type of classes newly analyzable through our extension of the PEP are function classes $\mathcal{C}_1$ and $\mathcal{C}_2$ of the form $g(Ax)$ and $h(x) + g(Ax)$. These three classes turn out to be included in $\mathcal{F}_{\mu,L}$ if we define the smoothness and strong convexity parameters of A, g and h in a proper way. Since the worst-case functions of $\mathcal{F}_{\mu,L}$ for (GM), found in Taylor et al. (2017), are sometimes but not always quadratic or of the form $g(Ax)$, we will quantify the performance gap between the general class $\mathcal{F}_{\mu,L}$ and the classes $\mathcal{Q}_{\mu,L}$, $\mathcal{C}_1$, $\mathcal{C}_2$ or other function classes involving matrices that we can now analyze through the PEP framework.

⋆ N. Bousselmi is supported by the French Community of Belgium through a FRIA fellowship (F.R.S-FNRS).

Theorems are stated without proofs in this extended abstract, they will be appear in a forthcoming paper.

2. PEP FORMULATION

Typically, a PEP can be formulated as follows. Given the class of functions $\mathcal{C}$, the optimization method $\mathcal{M}$ performing N iterations, the initial distance R and the classical performance criterion $f(x_N) - f^*$ (objective function accuracy after N iterations), the PEP is

$$\begin{aligned} \max_{x_0,\dots,x_N,f} \quad & f(x_N) - f^* \\ \text{s.t.} \quad & f \in \mathcal{C}, \\ & x_k \text{ generated by applying } \mathcal{M} \text{ to } f \text{ from } x_0, \\ & ||x_0 - x^*|| \leq R. \end{aligned} \tag{PEP}$$

We can study any method $\mathcal{M}$ that computes each iterate as a linear combination of the initial point x_0 and the gradients of the previous iterations, i.e.

$$x_k = x_0 - \sum_{i=0}^{k-1} h_{k,i}\nabla f(x_i).$$

Coefficients $h_{k,i}$ entirely describe the method $\mathcal{M}$. For example, the gradient method with constant step size $\frac{1}{L}$ started from x_0:

For $i = 0 : N-1$

$$\begin{aligned} x_{i+1} &= x_i - \frac{1}{L}\nabla f(x_i) \\ &= x_0 - \frac{1}{L}\sum_{i=0}^{k-1}\nabla f(x_i). \end{aligned} \tag{GM}$$

is described with

$$\begin{cases} h_{k,i} = \dfrac{1}{L} & \text{if } i < k, \\ h_{k,i} = 0 & \text{otherwise.} \end{cases}$$

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

The constraint $f \in \mathcal{C}$ must be expressed in an explicit way with *interpolation conditions* in order to have a tractable problem.

Definition 1. Given a set of triplet $\{(x_i, g_i, f_i)\}_{i\in I}$ with I some set of indices, *interpolation conditions* for the class of functions $\mathcal{C}$ are such that there exists a function $f \in \mathcal{C}$ with

$$\begin{aligned} f(x_i) &= f_i \quad \forall i \in I, \\ \nabla f(x_i) &= g_i \quad \forall i \in I, \end{aligned}$$

if and only if the *interpolation conditions* are satisfied.

When those conditions are available, the PEP can be rewritten as the following finite-dimensional problem

$$\begin{aligned} \max_{x_0,\dots,x_N,x_*,g_0,\dots,g_N,g_*,f_0,\dots,f_N,f_*} \quad & f_N - f_* \\ \text{s.t.} \quad & x_k = x_0 - \sum_{i=0}^{k-1} h_{k,i}\nabla g_i, \\ & ||x_0 - x_*||^2 \leq R^2, \\ & \|g_*\|^2 = 0, \\ & \{(x_i, g_i, f_i)\}_{i\in I=\{0,1,\dots,N,*\}} \text{ are interpolable} \\ & \text{by some function } f \in \mathcal{C}. \end{aligned} \tag{PEP}$$

Finally, it was shown in Taylor et al. (2017) that this problem becomes a convex semidefinite problem provided that the iterates x_i and their gradients g_i are represented as elements of the Gram matrix $G = P^T P$, with

$$P = \begin{pmatrix} x_1 & \cdots & x_N & g_1 & \cdots & g_N \end{pmatrix} \in \mathbb{R}^{d\times 2N}.$$

3. PROBLEM STATEMENT

The key step and our main contribution is to obtain interpolation conditions for the class $\mathcal{Q}_{\mu,L}$ of quadratic functions. Indeed, we want to solve the following PEP on the class $\mathcal{Q}_{\mu,L}$,

$$\begin{aligned} \max_{x_0,\dots,x_N,f} \quad & f(x_N) - f^* \\ \text{s.t.} \quad & f \in \mathcal{Q}_{\mu,L}, \\ & x_k \text{ generated by applying } \mathcal{M} \text{ to } f \text{ from } x_0, \\ & ||x_0 - x^*|| \leq R. \end{aligned} \tag{PEP-Q}$$

where we need an explicit equivalent reformulation of the condition $f \in \mathcal{Q}_{\mu,L}$ in order to solve (PEP-Q).

As mentioned above, (PEP) can be formulated under the form of a semidefinite program (see Taylor et al. (2017)) involving only the Gram matrix G of the iterates x_i and their gradients g_i and the values f_i of the function at these iterates.

In order to work in the class $\mathcal{Q}_{\mu,L}$, we must consider the set of Gram matrices associated to a quadratic function. Note that in that case we have

$$\nabla f(x) = Qx \quad \forall x \tag{1}$$

Definition 2. A symmetric matrix $G \in \mathbb{S}^{2N}$ is a (μ, L, N)-*quadratic-Gram matrix* if and only if there exist a dimension $d \in \mathbb{N}$, a symmetric matrix $Q \in \mathbb{S}^d$ with $\mu I \preceq Q \preceq LI$ and a sequence $x_i \in \mathbb{R}^d$ for $i = 1, \dots, N$ such that $G = P^T P$ with

$$P = \begin{pmatrix} x_1 & \cdots & x_N & \overbrace{Qx_1}^{g_1} & \cdots & \overbrace{Qx_N}^{g_N} \end{pmatrix} \in \mathbb{R}^{d\times 2N}.$$

The set of all (μ, L, N)-quadratic-Gram matrices is denoted $\mathcal{G}_{\mu,L,N}$. It can be shown that any conic combination of (μ, L, N)-quadratic-Gram matrices is also a (μ, L, N)-quadratic-Gram matrix, hence the set $\mathcal{G}_{\mu,L,N}$ is a convex cone.

In the following, we provide an explicit convex description of this set in order to be able to include those constraints to (PEP-Q). In other words, we show a convex formulation of the condition $f \in \mathcal{Q}_{\mu,L}$.

4. INTERPOLATION CONDITIONS

Several observations can be made about the form of the (μ, L, N)-quadratic-Gram matrices. Indeed, if a matrix G belongs to $\mathcal{G}_{\mu,L,N}$, then, by diagonalization of Q, it can be written under the form

$$\begin{aligned} G &= \begin{pmatrix} X^TX & X^TQX \\ X^TQX & X^TQ^2X \end{pmatrix} \\ &= \begin{pmatrix} Y^TY & Y^TDY \\ Y^TDY & Y^TD^2Y \end{pmatrix} \\ &= \sum_{k=1}^{d} \begin{pmatrix} u_k u_k^T & \lambda_k u_k u_k^T \\ \lambda_k u_k u_k^T & \lambda_k^2 u_k u_k^T \end{pmatrix} \end{aligned} \tag{2}$$

where $X = (x_1 \cdots x_N) \in \mathbb{R}^{d\times N}$, $Q = VDV^T$ is the eigenvalue decomposition of Q, $Y = V^TX \in \mathbb{R}^{d\times N}$, $D = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$, $\lambda_k \in [\mu, L]$ and $u_k = \begin{pmatrix} y_{1,k} \\ \vdots \\ y_{N,k} \end{pmatrix}$.

Vector u_k contains the k-th component of all vectors y_i. Expression (2) informs us that each block X^TX, X^TQX, X^TQ^2X can be expressed as the sum of d positive definite rank-1 matrices $u_k u_k^T$, $\lambda_k u_k u_k^T$ and $\lambda_k^2 u_k u_k^T$.

By characterizing the Gram matrices exhibiting this structure, we are able to obtain the following explicit description of (μ, L, N)-quadratic-Gram matrices.

Theorem 2. Given a symmetric matrix

$$G = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \in \mathbb{S}^{2N}$$

with $A, C \in \mathbb{S}^N$ and $B \in \mathbb{R}^{N\times N}$, the conditions

$$G \succeq 0 \tag{C1}$$

$$B = B^T \tag{C2}$$

$$B \succeq \frac{\mu L}{\mu + L} A + \frac{1}{\mu + L} C \tag{C3}$$

are necessary and sufficient for

$$G \in \mathcal{G}_{\mu,L,N}.$$

Observe that the quadratic interpolation conditions (C1), (C2) and (C3) of Theorem 2 do not involve the function values f_i. Actually, the variables f_i are directly encoded in the diagonal of the block matrix $B = X^TQX$. Indeed, thanks to (1), we have

$$f(x) = \frac{1}{2} x^T Q x = \frac{1}{2} x^T \nabla f(x)$$

and the iterates x_i, g_i and f_i are linked through

$$f_i = \frac{1}{2} x_i^T g_i.$$

Since B contains the scalar products $x_i^T g_j$, whenever we need the value of f_i, we just use

$$f_i = \frac{1}{2} B_{ii}.$$

It is now possible to replace the condition $f \in \mathcal{Q}_{\mu,L}$ in (PEP-Q) by the interpolation conditions obtained in Theorem 2, which allows to reformulate the whole problem as a tractable optimization problem. This problem is a convex semidefinite program involving linear matrix constraint and can be comfortably written and solved with the *Python* library `PEPit` (see Goujaud et al. (2022)).
Note that as we only consider homogeneous quadratic functions in the class $\mathcal{Q}_{\mu,L}$, we can assume implicitly that $x^* = 0$ and $f^* = f(x^*) = 0$, which simplifies the formulation.

Finally, as mentioned in the introduction, we actually obtained more general interpolation conditions that the ones of the class of quadratic functions. Indeed, Definition 2 and Theorem 2 provide interpolation conditions for any two sequences x_i and y_i linked by a matrix, i.e. $y_i = Qx_i \quad \forall i$. For example, if we apply a first-order method to the class of functions of the form $f(x) = g(Qx)$, then we will need to compute the gradient of f, i.e.

$$\nabla f(x) = Q \nabla g(Qx). \tag{3}$$

In order to describe this class with interpolation conditions, we need to force x_i and $y_i = Qx_i$ to be linked by a matrix as well as the $u_i = \nabla g(Qx_i)$ and $v_i = Q\nabla g(Qx_i)$. Thanks to Theorem 2, we are able to do it and, thus, to analyze the worst-case performance of this class through PEP.

5. RELATION BETWEEN INTERPOLATION CONDITIONS OF $\mathcal{F}_{\mu,L}$ AND $\mathcal{Q}_{\mu,L}$

The class of quadratic functions $\mathcal{Q}_{\mu,L}$ is included in $\mathcal{F}_{\mu,L}$, therefore, from the interpolation conditions of $\mathcal{Q}_{\mu,L}$, it must be possible to obtain the interpolation conditions of $\mathcal{F}_{\mu,L}$.

In Taylor et al. (2017), the following interpolation conditions for the class $\mathcal{F}_{\mu,L}$ have been obtained $\forall i, j = 0, 1, \dots, N$

$$\begin{aligned} f_i - f_j - g_j^T(x_i - x_j) \geq \frac{1}{2(1-\mu/L)} & \\ \left(\frac{1}{L}(g_i^T g_i + g_j^T g_j - 2 g_i^T g_j) \right. & \\ + \mu(x_i^T x_i + x_j^T x_j - 2 x_i^T x_j) & \\ \left. - 2\frac{\mu}{L}(g_j^T x_j - g_j^T x_i - g_i^T x_j + g_i^T x_i) \right). & \end{aligned} \tag{4}$$

In the quadratic case, if we define a matrix $M = -\frac{\mu L}{\mu+L} A + B - \frac{1}{\mu+L} C$, the condition (C3) can be written as the positive semidefiniteness of matrix M which is equivalent to

$$\begin{aligned} M \succeq 0 \Leftrightarrow z^T M z \geq 0 \quad \forall z \in \mathbb{R}^N & \\ \Leftrightarrow \sum_{k=1}^{N} \sum_{l=1}^{N} z_k z_l M_{kl} \geq 0 \quad \forall z \in \mathbb{R}^N. & \end{aligned} \tag{5}$$

Choosing $z_i = 1$, $z_j = -1$ and all the other components of z equal to zero in (5) and then using $f_i = \frac{1}{2} x_i^T g_i$ yields the interpolation conditions (4) of the class $\mathcal{F}_{\mu,L}$.

Therefore, the finite set of interpolation conditions of $\mathcal{F}_{\mu,L}$ is explicitly seen as a consequence of the set of interpolation conditions of $\mathcal{Q}_{\mu,L}$.

6. ANALYSIS OF THE GRADIENT METHOD

In Taylor et al. (2017), the worst-case performance and the functions reaching it for the class $\mathcal{F}_{\mu,L}$ have been completely analyzed thanks to the PEP methodology. We would like to compare these results with the behavior of (GM) on the class $\mathcal{Q}_{\mu,L}$ and other classes involving matrices.

In the convex case $\mu = 0$, the worst-case performance on the class $\mathcal{F}_{0,L}$ is (from Taylor et al. (2017))

$$f(x_N) - f^* \leq \frac{LR^2}{4N+2}. \tag{6}$$

Note that this worst-case performance is reached by a Huber function, which is not quadratic and does not belong to $\mathcal{Q}_{0,L}$.

Thanks to our extension of PEP for the class $\mathcal{Q}_{\mu,L}$, we can solve (PEP) for the class $\mathcal{Q}_{0,L}$. It yields the following numerical results. Fig. 1 is the worst-case performance of (GM) on $\mathcal{F}_{0,L}$ (red) and $\mathcal{Q}_{0,L}$ (blue) for each number of iterations N.

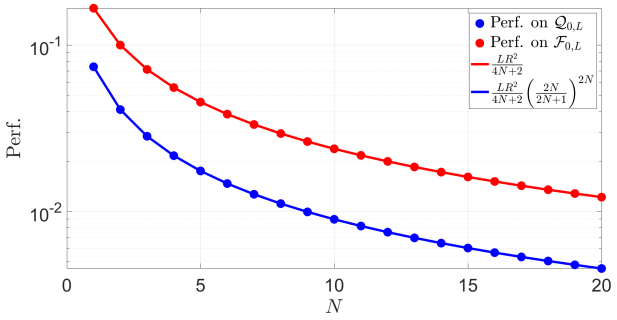


Fig. 1. Worst-case performance of (GM) on $\mathcal{Q}_{\mu,L}$ (blue dots) obtained by PEP and on $\mathcal{F}_{\mu,L}$ (red dots).

It turns out that it is possible to identify the worst-case rate of performance of (GM) on $\mathcal{Q}_{0,L}$, which is equal to the following analytical expression

$$f(x_N) - f^* \leq \frac{LR^2}{4N+2} \left(\frac{2N}{2N+1} \right)^{2N} \tag{7}$$

and this worst performance is achieved by the quadratic function

$$f(x) = \frac{Lx^2}{4N+2}.$$

We observe that the numerical results of PEP (blue dots) in Fig. 1 exactly matches the rate (7) (blue line).

Interestingly, the difference between the worst-case performance of (GM) on $\mathcal{F}_{0,L}$ and $\mathcal{Q}_{0,L}$ is the factor $\left(\frac{2N}{2N+1}\right)^{2N}$. Moreover, this factor exhibits two particular properties

$$\lim_{N\to\infty}\left(\frac{2N}{2N+1}\right)^{2N} = \frac{1}{e},$$
$$\left(\frac{2N}{2N+1}\right)^{2N} \geq \frac{1}{e} \quad \forall N \in \mathbb{N}.$$

Therefore, we can say that, for any number N of iterations, the worst-case performance of (GM) with constant step size $\frac{1}{L}$ on $\mathcal{F}_{0,L}$ is always lower than the performance on $\mathcal{Q}_{0,L}$ multiplied by a factor e.

To be complete, we must now mention that the literature already provides a methodology to analyze the worst-case performance of a first-order method on the class of quadratic functions with eigenvalues between μ and L (see for example Flanders and Shortley (1950); Nemirovsky and Polyak (1984); d'Aspremont et al. (2021)) and, thus, to obtain the rate (7). Indeed, given a quadratic function $\frac{1}{2}x^TQx$, an initial point x_0 and a method $\mathcal{M}$, the maximization of the last iterate x_N can be expressed as the maximization of a polynomial evaluated at the elements of the spectrum of Q, where the coefficients of the polynomial only depend on the method $\mathcal{M}$. Therefore this leads to the maximization of some explicit polynomial whose degree grows with the number of iterations. It can be shown that such a reasoning will provide the same rate (7).

However, as explained earlier, we are now also able to analyze the class of functions of the form $g(Ax)$, which cannot be tackled by the simple polynomial approach described in the previous paragraph. We observe a difference of worst-case performance of (GM) between the general class $\mathcal{F}_{0,L}$ and the class $\mathcal{C}_1$ of functions of the form $f(x) = g(Ax)$ where f is still L-smooth convex. Fig. 2 is the worst-case performance of (GM) on $\mathcal{F}_{0,L}$ (red) and $\mathcal{C}_1$ (blue) for each number of iterations N. Note again that such

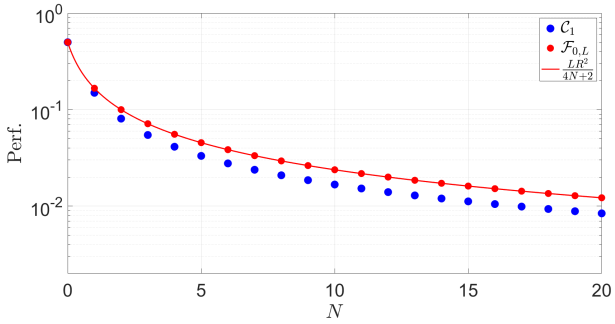


Fig. 2. Worst-case performance of (GM) on C_1 (blue dots) obtained by PEP.

results cannot be obtained by the abovementioned spectral analysis, and that it is possible with PEP to study even more complex functions classes such as $h(x) + g(Ax)$.

7. CONCLUSION

PEP has been shown to be a powerful tool for the analysis of the worst-case behavior of first-order optimization methods on a given class of functions. We showed how to extend PEP to the class of quadratic functions, thanks to Theorem 2, using a list of explicit convex constraints on the Gram matrix G. Moreover, we are able to implement and solve the PEP thanks to the *Python* library `PEPit`.

Our numerical experiments exactly match the analytical expression of the worst-case performance of the gradient method on convex quadratic functions $\mathcal{Q}_{0,L}$ and we compared it to the worst-case performance on smooth strongly convex functions $\mathcal{F}_{0,L}$. An interesting direction for future research would be to obtain a bound on the performance gap of any method between the general class $\mathcal{F}_{\mu,L}$ and the class $\mathcal{Q}_{\mu,L}$.

Moreover, in addition to the class of quadratic functions, we are now able to formulate explicit interpolation conditions for any class of functions involving matrices and to analyze them through PEP. This include for example the simple class of functions $f(x) = g(Ax)$ but also more complicated classes of functions as $f(x) = h(x) + g(Ax)$. Although the worst-case performance on the class of quadratic methods could already be obtained via the spectrum analysis approach, our extension of PEP appears to the best of our knowledge to be the first tool able to analyze classes of function of the forms $f(x) = g(Ax)$ or $f(x) = h(x) + g(Ax)$.

REFERENCES

A.B. Taylor, J.M. Hendrickx, F. Glineur *Smooth strongly convex interpolation and exact worst-case performance of first-order methods.* Mathematical Programming, 2017, vol. 161, no 1, p. 307-345.

B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, A. Dieuleveut (2022). *PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python.*

Y. Drori, M. Teboulle *Performance of first-order methods for smooth convex minimization: a novel approach.* Mathematical Programming, 2014, vol. 145, no 1, p. 451-482.

D. Scieur, F. Pedregosa *Universal average-case optimality of Polyak momentum.* International Conference on Machine Learning, 2020, (pp. 8565-8572). PMLR.

A. d'Aspremont, D. Scieur, A. Taylor *Acceleration methods.* Foundations and Trends® in Optimization, 2021, vol. 5, no 1-2, p. 1-245.

D. A. Flanders, G. Shortley *Numerical determination of fundamental modes.* Journal of Applied Physics. 21(12): 1326–1332.

A. S. Nemirovsky, B. T. Polyak *Iterative methods for solving linear ill-posed problems under precise information.* ENG. CYBER. (4): 50–56.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A Hankel realization for noncommutative rational functions around a matrix point

Motke Porat [*] **Victor Vinnikov** [**]

[*] *Dept. Math., Ben Gurion Univ. of the Negev, Beer-Sheva, Israel*
(e-mail: motpor@gmail.com*)*
[**] *Dept. Math., Ben Gurion Univ. of the Negev, Beer-Sheva, Israel*
(e-mail: vinnikov@math.bgu.ac.il*)*

Abstract: It is well known that noncommutative (nc) rational functions regular at the origin admit a good realization (or linearization) theory. This is very useful both conceptually and for a variety of applications since it often essentially reduces the study of these rational functions to a study of linear pencils. By translation the method can be applied to nc rational functions that are regular at some scalar point, but not beyond. In this talk we discuss the realization problem for nc rational functions regular at an arbitrary given matrix point using the nc difference–differential calculus and the general Taylor–Taylor series of nc function theory Kaliuzhnyi-Verbovetskyi and Vinnikov (2014) and provide a solution which is the analogue of the classical Hankel realization.

Keywords: Noncommutative Rational Functions, Free Noncommutative Function Theory, Taylor–Taylor Series, Noncommutative Multidimensional Systems

NC RATIONAL FUNCTIONS

We consider the ring of nc polynomials (the free ring) $\mathbb{K}\langle x_1, \dots, x_d\rangle$ over a field $\mathbb{K}$. Here $x_1, \dots, x_d$ are nc indeterminates, and $p \in \mathbb{K}\langle x_1, \dots, x_d\rangle$ is of the form

$$p = \sum_{w \in \boldsymbol{\mathcal{G}}_d} p_w x^w, \tag{1}$$

where $\boldsymbol{\mathcal{G}}_d$ denotes the free monoid on d generators (letters) $g_1, \dots, g_d$ with identity $\emptyset$ (the empty word), $p_w \in \mathbb{K}$, x^w are nc monomials in $x_1, \dots, x_d$ ($x^w = x_{j_1} \cdots x_{j_l}$ for $w = g_{j_1} \cdots g_{j_l} \in \boldsymbol{\mathcal{G}}_d$ and $x^\emptyset = 1$), and the sum is finite. p can be evaluated in an obvious way on d-tuples of square matrices of all sizes over $\mathbb{K}$: for $X = (X_1, \dots, X_d) \in (\mathbb{K}^{n\times n})^d$,

$$p(X) = \sum_{w \in \boldsymbol{\mathcal{G}}_d} p_w X^w = \sum_{w \in \boldsymbol{\mathcal{G}}_d} X^w p_w \in \mathbb{K}^{n\times n}. \tag{2}$$

Notice that nonzero p can vanish on $(\mathbb{K}^{n\times n})^d$ for some n: $p = x_1x_2 - x_2x_1$ vanishes on $\left(\mathbb{K}^{1\times 1}\right)^d$, and $p = \sum_{\pi \in \mathcal{S}_{n+1}} \operatorname{sign}(\pi)\, x_1^{\pi(1)-1} x_2 \cdots x_1^{\pi(n+1)-1} x_2$ (where $\mathcal{S}_{n+1}$ is the symmetric group on $n+1$ elements) vanishes on $(\mathbb{K}^{n\times n})^d$. However if $p(X) = 0$ for all $X \in \coprod_{n=1}^\infty (\mathbb{K}^{n\times n})^d$ then $p = 0$.

The skew field of nc rational functions $\mathbb{K}\langle\langle x_1, \dots, x_d\rangle\rangle$ over a field $\mathbb{K}$ (the free skew field) is the universal skew field of fractions of the ring of nc polynomials over $\mathbb{K}$. This involves some non-trivial details since unlike the commutative case, a nc rational function does not admit a canonical coprime fraction representation; see Amitsur (1966); Bergman (1970); Cohn (1971a, 1972) for some of the original constructions, and Rowen (1980) (Chapter 8) and Cohn (1971b, 2006) for good expositions and background. The following is most natural from the point of view of nc function theory and is a version of Amitsur's original construction except that we use evaluation on d-tuples of square matrices of all sizes over $\mathbb{K}$ instead of evaluation on a "large" auxilliary skew field; see Kaliuzhnyi-Verbovetskyi and Vinnikov (2009, 2012) for details and further references. We first define (scalar) nc rational expressions by starting with nc polynomials and then applying successive arithmetic operations — addition, multiplication, and inversion. A nc rational expression r can be evaluated on a d-tuple X of $n\times n$ matrices in its *domain of regularity*, $\operatorname{dom} r$, which is defined as the set of all d-tuples of square matrices of all sizes such that all the inverses involved in the calculation of $r(X)$ exist. (We assume that $\operatorname{dom} r \neq \emptyset$, in other words, when forming nc rational expressions we never invert an expression that is nowhere invertible.) Two nc rational expressions r_1 and r_2 are called *equivalent* if $\operatorname{dom} r_1 \cap \operatorname{dom} r_2 \neq \emptyset$ and $r_1(Z) = r_2(Z)$ for all d-tuples $Z \in \operatorname{dom} r_1 \cap \operatorname{dom} r_2$. We define a *nc rational function* $\mathfrak{r}$ to be an equivalence class of nc rational expressions; notice that it has a well-defined evaluation on $\operatorname{dom} \mathfrak{r} = \bigcup_{R\in\mathfrak{r}} \operatorname{dom} R$ (here R denotes a 1×1 matrix-valued rather than scalar nc rational expression, i.e., some of the intermediate expressions may involve matrices of scalar nc rational expressions, cf. below). We set $(\operatorname{dom}\mathfrak{r})_n = \operatorname{dom}\mathfrak{r} \cap (\mathbb{K}^{n\times n})^d$.

It is clear that the evaluation of a nc rational function respects direct sums and simultaneous similarities, so that a nc rational function $\mathfrak{r}$ defines a nc function (Kaliuzhnyi-Verbovetskyi and Vinnikov (2014)) on $\operatorname{dom}\mathfrak{r}$ (technically, on an apriori somewhat larger set called the extended domain of regularity of $\mathfrak{r}$ obtained by evaluating $\mathfrak{r}$ on d-tuples of generic matrices). In particular, nc rational functions admit a difference-differential calculus. The partial nc difference-differential operators

$$\Delta_j \colon \mathbb{K}\langle\langle x_1, \dots, x_d\rangle\rangle \to \mathbb{K}\langle\langle x_1, \dots, x_d\rangle\rangle \otimes \mathbb{K}\langle\langle x_1, \dots, x_d\rangle\rangle,$$

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

$j = 1, \dots, d$, can be defined recursively (on the level of nc rational expressions) starting with $\Delta_j(x_i) = \delta_{ij}(1 \otimes 1)$ and using the nc calculus rules, see Kaliuzhnyi-Verbovetskyi and Vinnikov (2012).

POWER SERIES EXPANSION AROUND A SCALAR POINT

A nc rational expression which is regular at 0 determines a nc formal power series. This correspondence is defined recursively using addition and multiplication of nc formal power series and inversion of a nc formal power series with an invertible constant term (the coefficient of $z^{\emptyset}$). Furthermore, r_1 and r_2 are equivalent if and only if the corresponding nc formal power series coincide. By translation, if $\lambda = (\lambda_1, \dots, \lambda_d) \in (\operatorname{dom} \mathfrak{r})_1 \subseteq \mathbb{K}^d$ we obtain, for $X = (X_1, \dots, X_d) \in \mathbb{K}^{n \times n}$,

$$\mathfrak{r}(X) \sim \sum_{w \in \boldsymbol{\mathcal{G}}_d} (X - I_n \lambda)^w \mathfrak{r}_w. \tag{3}$$

Here $\mathfrak{r}_w \in \mathbb{K}$ are the coefficients, and $X - I_n\lambda$ stands for $(X_1 - I_n\lambda_1, \dots, X_d - I_n\lambda_d)$.

From the point of view of nc function theory, (3) is the Taylor–Taylor (TT) power series expansion of $\mathfrak{r}$ around λ. In particular, the coefficients $\mathfrak{r}_w$ can be calculated by means of the nc difference-differential calculus: $\mathfrak{r}_w = \Delta^w \mathfrak{r}(\lambda, \dots, \lambda)$, Also, the series (3) actually converges to $\mathfrak{r}(X)$ in the following cases: (a) if $X - I_n\lambda$ is a jointly nilpotent d tuple of matrices so that the sum is finite; (b) in the case $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, the series converges normally on any open nc ball $\coprod_{n=1}^{\infty} \left\{ X \in (\mathbb{K}^{n \times n})^d : \|X - I_n\lambda\| < r \right\}$ (with respect to any operator space norm on $\mathbb{K}^d$, e.g., $\|Z\| = \|Z_1^* Z_1 + \cdots + Z_d^* Z_d\|$ for $Z = (Z_1, \dots, Z_d)$) contained in the (extended) domain of regularity of $\mathfrak{r}$.

REALIZATION THEORY AROUND THE ORIGIN

We have the following basic facts of life (Kleene (1956), Schützenberger (1961, 1963), Fliess (1970, 1974a,b), see Berstel and Reutenauer (1988) for a good survey, Ball, Groenewald, and Malakorn (2005, 2006a,b), Kaliuzhnyi-Verbovetskyi and Vinnikov (2009, 2012)):

(1) A nc power series $\sum_{w \in \boldsymbol{\mathcal{G}}_d} \mathfrak{r}_w x^w \in \mathbb{K}\langle\langle x_1, \dots, x_d \rangle\rangle$ is the power series expansion of a nc rational function at a scalar point iff the corresponding infinite $\boldsymbol{\mathcal{G}}_d \times \boldsymbol{\mathcal{G}}_d$ *Hankel matrix* $\mathbb{H} = [\mathfrak{r}_{uv}]_{u,v \in \boldsymbol{\mathcal{G}}_d}$ has finite rank.

(2) If a nc rational function $\mathfrak{r}$ is regular at 0 it admits a unique (up to unique similarity) minimal (controllable and observable) state space realization:

$$\mathfrak{r}(x) = D + C(I - A_1 x_1 - \cdots - A_d x_d)^{-1} (B_1 x_1 + \cdots + B_d x_d), \tag{4}$$

where $A_1, \dots, A_d \in \mathbb{K}^{L \times L}$ for some integer L, $B_1, \dots, B_d \in \mathbb{K}^{L \times 1}$, $C \in \mathbb{K}^{1 \times L}$, and $D = \mathfrak{r}(0)$. Furthermore,

$$\operatorname{dom} \mathfrak{r} = \coprod_{n=1}^{\infty} \left\{ X = (X_1, \dots, X_d) \in \left(\mathbb{K}^{n \times n}\right)^d : \det(I_{Ln} - X_1 \otimes A_1 - \cdots - X_d \otimes A_d) \neq 0 \right\}.$$

Here a realization is called *minimal* if the state space dimension L is as small as possible, *controllable* if

$$\operatorname{span}_{i=1,\dots,d,\, w \in \boldsymbol{\mathcal{G}}_d} \operatorname{im} A^w B_i = \mathbb{C}^L,$$

and *observable* if

$$\cap_{w \in \boldsymbol{\mathcal{G}}_d} \ker C A^w = \{0\}.$$

The items (1) and (2) are closely related: the realization (4) implies immediately that the column rank of the Hankel matrix $\mathbb{H}$ is at most L, and whereas a realization can be constructed recursively by synthesis, starting with polynomuals (or even just the basic monomials $x_1, \dots, x_d$) and using sum, product, and inversion formulae, it can also be constructed in one step using the columns space of $\mathbb{H}$. We refer to Ball, Groenewald, and Malakorn (2005) and Kaliuzhnyi-Verbovetskyi and Vinnikov (2012) for details and further references.

POWER SERIES EXPANSION AROUND A MATRIX POINT

Some notation: for $P = [P_{ij}]_{i,j=1,\dots,m}, Q = [Q_{ij}]_{i,j=1,\dots,m} \in \mathbb{K}^{sm \times sm} \cong \left(\mathbb{K}^{s \times s}\right)^{m \times m}$, we let $P \odot_s Q$ denote the product of P and Q viewed as $m \times m$ matrices over the tensor algebra of $\mathbb{K}^{s \times s}$:

$$P \odot_s Q = \left[\sum_{j=1}^{m} P_{ij} \otimes Q_{jk} \right]_{i,k=1,\dots,m} \in \left(\mathbb{K}^{s \times s} \otimes \mathbb{K}^{s \times s}\right)^{m \times m}.$$

For $Z = (Z_1, \dots, Z_d) \in \left(\mathbb{K}^{sm \times sm}\right)^d$ and $w = g_{j_1} \cdots g_{j_l} \in \boldsymbol{\mathcal{G}}_d$, we let $Z^{\odot_s w} = Z_{i_1} \odot_s \cdots \odot_s Z_{i_l} \in \left(\left(\mathbb{K}^{s \times s}\right)^{\otimes l}\right)^{m \times m}$.

The power series expansion around $Y \in (\operatorname{dom} \mathfrak{r})_s$ is now given by, for $X \in \left(\mathbb{K}^{sm \times sm}\right)^d$,

$$\mathfrak{r}(X) \sim \sum_{w \in \boldsymbol{\mathcal{G}}_d} (X - I_m \otimes Y)^{\odot_s w}\, \mathfrak{r}_w. \tag{5}$$

Here, the coefficient $\mathfrak{r}_w$ is a l-linear mapping $\left(\mathbb{K}^{s \times s}\right)^l \longrightarrow \mathbb{K}^{s \times s}$, where l is the length of the word w, or alternatively a linear mapping $\left(\mathbb{K}^{s \times s}\right)^{\otimes l} \longrightarrow \mathbb{K}^{s \times s}$. Notice that

$$\left(X - \bigoplus_{\alpha=1}^{m} Y \right)^{\odot_s w} \in \left(\left(\mathbb{K}^{s \times s}\right)^{\otimes l}\right)^{m \times m},$$

hence we can apply $\mathfrak{r}_w$ to every entry of this matrix yielding a matrix in $\left(\mathbb{K}^{s \times s}\right)^{m \times m} \cong \mathbb{K}^{sm \times sm}$ — which is where the value $\mathfrak{r}(X)$ lies.

NC formal power series with a matrix centre Y, of the form (5), form a ring with an obvious convolution product. It is clear that any nc polynomial can be written as a (finite) nc power series with centre Y, and the power series expansion of a nc rational expression r regular at Y can be obtained recursively using addition and multiplication of nc formal power series with centre Y and inversion of a nc formal power series with an invertible constant term. From the point of view of nc function theory, (3) is the TT power series expansion of $\mathfrak{r}$ around a matrix centre Y. One important difference with the case of a scalar centre is that the coefficients $\mathfrak{r}_w$ are not arbitrary multilinear mappings; they have to satisfy certain compatibility conditions with respect to Y, see Kaliuzhnyi-Verbovetskyi and Vinnikov (2014) ((4.14)–(4.17)).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

REALIZATION THEORY AROUND A MATRIX POINT

We have shown in Porat and Vinnikov (2021, 2020) that if a nc rational function $\mathfrak{r}$ is regular at $Y \in (\mathbb{K}^{s\times s})^d$, it admits a unique (up to unique similarity) minimal (controllable and observable) state space realization with centre Y: for $X \in (\mathbb{K}^{sm\times sm})^d$,

$$\mathfrak{r}(X) = I_m \otimes D + (I_m\otimes C)\big(I_{Lsm} - (X_1 - I_m\otimes Y_1)\mathbf{A}_1 - \cdots - (X_d - I_m\otimes Y_d)\mathbf{A}_d\big)^{-1} \big((X_1 - I_m\otimes Y_1)\mathbf{B}_1 + \cdots + (X_d - I_m\otimes Y_d)\mathbf{B}_d\big). \quad (6)$$

Here $\mathbf{A}_1, \ldots, \mathbf{A}_d \colon \mathbb{K}^{s\times s} \to \mathbb{K}^{Ls\times Ls}$ for some integer L and $\mathbf{B}_1, \ldots, \mathbf{B}_d \colon \mathbb{K}^{s\times s} \to \mathbb{K}^{Ls\times s}$ are linear mappings, $C \in \mathbb{K}^{s\times Ls}$, and $D = \mathfrak{r}(Y) \in \mathbb{K}^{s\times s}$. Furthermore,

$$(\operatorname{dom}\mathfrak{r})_{sm} = \big\{X = (X_1, \ldots, X_d) \in \big(\mathbb{K}^{sm\times sm}\big)^d : \det\big(I_{Lsm} - (X_1 - I_m\otimes Y_1)\mathbf{A}_1 - \cdots - (X_d - I_m\otimes Y_d)\mathbf{A}_d\big) \neq 0\big\}.$$

In contrast to the case of realizations around the origin (or around a scalar centre), there are necessary and sufficient compatibility conditions with respect to Y on the coefficients $\mathbf{A}_1, \ldots, \mathbf{A}_d$, $\mathbf{B}_1, \ldots, \mathbf{B}_d$, C, D for the corresponding sequence of coefficients $\mathfrak{r}_w$ to satisfy the compatibility conditions of Kaliuzhnyi-Verbovetskyi and Vinnikov (2014) ((4.14)–(4.17)) and for (6) to define a nc rational function.

HANKEL REALIZATIONS AND A GENERALIZED FLIESS–KRONECKER THEOREM

The construction of the realisation (6) in Porat and Vinnikov (2021, 2020) involves synthesis using sum, production, and inversion formulae. In this talk we will use the nc difference–differential calculus to associate to the realisation (6) a functional model that is obtained directly from the function $\mathfrak{r}$ and the $s \times s$ matrix centre Y. This leads to a generalization of the Fliess–Kronecker theorem. Let $\mathbb{I} = \bigcup_{\ell=0}^{\infty} \mathbb{I}_\ell$, $\mathbb{I}_\ell = \{\omega \in \boldsymbol{\mathcal{G}}_d \colon |\omega| = \ell\} \times (\{1, \ldots, s\}\times\{1, \ldots, s\})^\ell$. Then the nc power series (5) with centre Y, with the coefficients satisfying the corresponding compatibility conditions of Kaliuzhnyi-Verbovetskyi and Vinnikov (2014) ((4.14)–(4.17)) with respect to Y, is the power series expansion at Y of a nc rational function iff the infinite Hankel matrix

$$\mathbb{H} = \Big[\mathfrak{r}_{uu'}\Big(E_{i_1,j_1}, \ldots, E_{i_{|u|},j_{|u|}}, E_{i'_1,j'_1}, E_{i'_{|u'|},j'_{|u'|}}\Big)\Big]$$

where $(u, (i_1, j_1), \ldots, (i_{|u|}, j_{|u|}))$, $(u', (i'_1, j'_1), \ldots, (i'_{|u'|}, j'_{|u'|})) \in \mathbb{I}$ has finite rank.

The power series expansions and realizations around an arbitrary matrix point provide a direct construction of the free skew field as the limit of the corresponding local rings.

REFERENCES

S.A. Amitsur, *Rational identities and applications to algebra and geometry*, J. Algebra **3** (1966), 304–359.

J.A. Ball, G. Groenewald, and T. Malakorn, *Structured noncommutative multidimensional linear systems*, SIAM J. Control Optim. **44** (2005), 1474–1528.

J.A. Ball, G. Groenewald, and T. Malakorn, *Conservative structured noncommutative multidimensional linear systems*, in The State Space Method, Generalizations and Applications, Oper. Theory Adv. Appl. **161**, pp. 179–223, Birkhäuser, Basel (2006).

J.A. Ball, G. Groenewald, and T. Malakorn, *Bounded real lemma for structured noncommutative multidimensional linear systems and robust control*, Multidimens. Syst. Signal Process. **17** (2006), 119–150.

G.M. Bergman, *Skew fields of noncommutative rational functions, after Amitsur*, in Séminaire Schützenberger–Lentin–Nivat **16** (Année 1969/70), Paris (1970).

J. Berstel and C. Reutenauer, *Rational series and their languages*, EATCS Monographs on Theoretical Computer Science **12**, Springer-Verlag, Berlin (1988).

P.M. Cohn, *The embedding of firs in skew fields*, Proc. London Math. Soc. **23** (1971), 193–213.

P.M. Cohn, *Free rings and their relations*, London Mathematical Society Monographs **2**, Academic Press, London (1971).

P.M. Cohn, *Universal skew fields of fractions*, Symposia Math. **8** (1972), 135–148.

P.M. Cohn, *Free ideal rings and localization in general rings*, New Mathematical Monographs **3**, Cambridge University Press, Cambridge (2006).

M. Fliess, *Sur le plongement de l'algèbre des séries rationnelles non commutatives dans un corps gauche*, C. R. Acad. Sci. Paris, Ser. A **271** (1970), 926–927.

M. Fliess, *Matrices de Hankel*, J. Math. Pures Appl. **53** (1974), 197–222.

M. Fliess, *Sur divers produits de séries formelles*, Bull. Soc. Math. France **102** (1974), 181–191.

D.S. Kaliuzhnyi-Verbovetskyi and V. Vinnikov, *Singularities of Noncommutative Rational Functions and Minimal Factorizations*, Lin. Alg. Appl. **430** (2009), 869–889.

D.S. Kaliuzhnyi-Verbovetskyi and V. Vinnikov, *Noncommutative rational functions, their difference-differential calculus and realizations*, Multidimens. Syst. Signal Process. **23** (2012), 49–77.

D.S. Kaliuzhnyi-Verbovetzkyi and V. Vinnikov, *Foundations of Free Noncommutative Function Theory*, Math. Surveys and Monographs **199**, Amer. Math. Society (2014).

S.C. Kleene, *Representation of events in nerve nets and finite automata*, in Automata Studies, Annals of Mathematics Studies **34**, pp. 3–41, Princeton University Press, Princeton, N. J. (1956).

M. Porat, V. Vinnikov, Realizations of non-commutative rational functions around a matrix centre, I: synthesis, minimal realizations and evaluation on stably finite algebras, *J. London Math. Soc.* 104 (2021), 1250–1299.

M. Porat, V. Vinnikov, Realizations of non-commutative rational functions around a matrix centre, II: The lost-abbey conditions, preprint `arXiv:2009.08527`.

L.H. Rowen, *Polynomial identities in ring theory*, Pure and Applied Mathematics **84**, Academic Press Inc. (Harcourt Brace Jovanovich Publishers), New York (1980).

M.P. Schützenberger, *On the definition of a family of automata*, Information and Control **4** (1961), 245–270.

M.P. Schützenberger, *Certain elementary families of automata*, in Proc. Sympos. Math. Theory of Automata (New York, 1962), pp. 139–153, Polytechnic Press of Polytechnic Inst. of Brooklyn, Brooklyn, New York (1963).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A Hankel realization for discrete-time overdetermined systems

Joseph A. Ball * **Victor Vinnikov** **

* *Department of Mathematics, Virginia Tech, Blacksburg, VA USA (e-mail:* joball@math.vt.edu*)*
** *Dept. Math., Ben Gurion Univ. of the Negev, Beer-Sheva, Israel (e-mail:* vinnikov@math.bgu.ac.il*)*

Abstract: We consider overdetermined multidimensional discrete-time systems where the evolution of the whole state vector is given by several update equations in several linearly independent directions. Such systems are overdetermined and we assume that they come equipped with compatibility difference equations for the input and output signals. As a consequence of these compatibility equations frequency domain analysis leads to function theory on a certain algebraic curve rather than to function theory in several complex variables. More precisely, the transfer function of the system is (under certain assumptions) a meromorphic bundle map on a compact Riemann surface. In this talk we will discuss the corresponding realization problem and provide a solution which is the higher genus analogue of the classical Hankel realization.

Keywords: Multidimensional Systems, Algebraic Curves, Vector Bundles on a Compact Riemann Surface

OVERDETERMINED SYSTEMS AND COMPATIBILITY DIFFERENCE EQUATIONS

We consider a 2D (linear, time-invariant) discrete-time input/state/output (i/s/o) system given by

$$\Sigma: \quad \begin{cases} x(t+\mathbf{e}_1) &= A_1 x(t) + B_1 u(t) \\ x(t+\mathbf{e}_2) &= A_2 x(t) + B_2 u(t) \\ y(t) &= Cx(t) + Du(t) \end{cases} \tag{1}$$

where $t \in \mathbb{Z}^2$, $u(t)$ takes values in the *input space* $\mathcal{E}$, $x(t)$ takes values in the *state space* $\mathcal{X}$, and $y(t)$ takes values in the *output space* $\mathcal{E}_*$ and $\mathbf{e}_1$, $\mathbf{e}_2$ denote the two standard basis vectors

$$\mathbf{e}_1 = (1,0), \quad \mathbf{e}_2 = (0,1)$$

for $\mathbb{R}^2$, here used to indicated increments in the horizontal and vertical directions, respectively, on the integer lattice $\mathbb{Z}$. The system is *overdetermined* since there are two ways to compute $x(t+\mathbf{e}_1+\mathbf{e}_2)$ from $x(t)$:

$$x(t) \mapsto x(t+\mathbf{e}_1) \mapsto x(t+\mathbf{e}_1+\mathbf{e}_2),$$
$$x(t) \mapsto x(t+\mathbf{e}_2) \mapsto x(t+\mathbf{e}_2+\mathbf{e}_1).$$

This leads to the necessity of consideration of *compatibility conditions* to be satisfied in order that the system equations have solutions.

Requiring the system equations to be compatible for the free evolution (zero input) and arbitrary initial state leads to the commutativity condition:

$$A_1 A_2 = A_2 A_1. \tag{V1}$$

Analysis of non-free evolution leads us to assume that we have factorizations

$$B_1 = \widetilde{B}\sigma_1, \quad B_2 = \widetilde{B}\sigma_2,$$
$$A_2 B_1 - A_1 B_2 = A_2\widetilde{B}\sigma_1 - A_1\widetilde{B}\sigma_2 = \widetilde{B}\gamma \tag{V2}$$

for operators σ_1, σ_2 and γ mapping the input space $\mathcal{E}$ into some auxiliary input space $\widetilde{\mathcal{E}}$ and an input operator $\widetilde{B}\colon \widetilde{\mathcal{E}} \to \mathcal{X}$; then a sufficient condition for the input signal to be compatible is given by the input difference equation:

$$\sigma_2 u(t+\mathbf{e}_1) - \sigma_1 u(t+\mathbf{e}_2) + \gamma u(t) = 0. \tag{2}$$

We seek a similar output difference equation for the corresponding output signal

$$\sigma_{2*} y(t+\mathbf{e}_1) - \sigma_{1*} y(t+\mathbf{e}_2) + \gamma_* y(t) = 0 \tag{3}$$

for some σ_{1*}, σ_{2*} and γ_* in $\mathcal{L}(\mathcal{E}_*\widetilde{\mathcal{E}}_*)$. Using the system equations, this leads us naturally to assume that there is an operator $\widetilde{D} \in \mathcal{L}(\widetilde{\mathcal{E}}, \widetilde{\mathcal{E}}_*)$ so that

$$\sigma_{1*} C A_2 - \sigma_{2*} C A_1 = \gamma_* C, \tag{V3}$$

$$\sigma_{2*} D = \widetilde{D}\sigma_2, \quad \sigma_{1*} D = \widetilde{D}\sigma_1,$$
$$\sigma_{2*} C\widetilde{B}\sigma_1 - \sigma_{1*} C\widetilde{B}\sigma_2 + \gamma_* D = \widetilde{D}\gamma. \tag{V4}$$

A collection of spaces and operators satisfying (V1)–(V4) is called a (commutative two-operator) vessel. It corresponds to the overdetermined system (1) with the compatibility difference equations (2) & (3), as well as to the dual system

$$\Sigma_*: \quad \begin{cases} x_*(t-\mathbf{e}_1) &= A_1^* x_*(t) + C^*\sigma_{1*}^* u_*(t) \\ x_*(t-\mathbf{e}_2) &= A_2^* x_*(t) + C^*\sigma_{1*}^* u_*(t) \\ y_*(t) &= \widetilde{B}^* x_*(t) + \widetilde{D}^* u_*(t) \end{cases} \tag{4}$$

with state space $\mathcal{X}$, input space $\widetilde{\mathcal{E}}_*$ and output space $\widetilde{\mathcal{E}}$, and compatibility difference equations

$$\sigma_{2*}^* u_*(t-\mathbf{e}_1) - \sigma_{1*}^* u_* t - \mathbf{e}_2) + \gamma_*^* u_*(t) = 0 \tag{5}$$

and

$$\sigma_2^* y_*(t-\mathbf{e}_1) - \sigma_1^* y_*(t-\mathbf{e}_2) + \gamma^* y_*(t) = 0. \tag{6}$$

These systems are the discrete-time analogues of overdetermined 2D continuous-time systems that were discovered

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

by Livšic (1978, 1979a,b). They were extensively studied, see Livšic–Kravitsky–Markus–Vinnikov (1995); Vinnikov (1998); Livšic (2001); Ball and Vinnikov (2003), primarily as a system-theoretic tool for spectral analysis of pairs of commuting nonselfadjoint (primarily dissipative) operators [1], and for applying the state-space method to function theory on algebraic curves and compact Riemann surfaces. Assorted applications were considered in Livšic (1997, 2002).

FREQUENCY DOMAIN ANALYSIS

We consider a discrete wave trajectory

$$u(t) = \lambda_1^{t_1}\lambda_2^{t_2}u_0,\ x(t) = \lambda_1^{t_1}\lambda_2^{t_2}x_0,\ y(t) = \lambda_1^{t_1}\lambda_2^{t_2}y_0,$$

$t = (t_1, t_2) \in \mathbb{Z}^2$, where $\lambda = (\lambda_1, \lambda_2) \in \mathbb{C}^2$ (double frequency), $u_0 \in \mathcal{E}$, $x_0 \in \mathcal{X}$, $y_0 \in \mathcal{E}_*$ (amplitudes). We assume that the four input and output spaces are all finite dimensional and that (non-degeneracy assumption) $\exists \xi_1, \xi_2$ s.t. $\xi_1\sigma_1 + \xi_2\sigma_2$, $\xi_1\sigma_{1*} + \xi_2\sigma_{2*}$ are invertible. It follows that $\dim \mathcal{E} = \dim \widetilde{\mathcal{E}}$, $\dim \mathcal{E}_* = \dim \widetilde{\mathcal{E}}_*$, and we define the input discriminant polynomial, the input discriminant curve, and the input family of subspaces threreupon:

$$\begin{aligned}
&\mathbf{p}(\lambda) = \det(\lambda_1\sigma_2 - \lambda_2\sigma_1 + \gamma),\\
&\mathbf{C}_0 = \left\{\lambda \in \mathbb{C}^2 \colon \mathbf{p}(\lambda) = 0\right\},\\
&\boldsymbol{\mathcal{E}}(\lambda) = \ker(\lambda_1\sigma_2 - \lambda_2\sigma_1 + \gamma) \text{ for } \lambda \in \mathbf{C}_0,
\end{aligned}$$

and similarly on the output side:

$$\begin{aligned}
&\mathbf{p}_*(\lambda) = \det(\lambda_1\sigma_{2*} - \lambda_2\sigma_{1*} + \gamma_*),\\
&\mathbf{C}_{0*} = \left\{\lambda \in \mathbb{C}^2 \colon \mathbf{p}_*(\lambda) = 0\right\},\\
&\boldsymbol{\mathcal{E}}_*(\lambda) = \ker(\lambda_1\sigma_{2*} - \lambda_2\sigma_{1*} + \gamma_*) \text{ for } \lambda \in \mathbf{C}_{0*}.
\end{aligned}$$

We conclude that if we are given a wave trajectory as above with $u_0 \neq 0$, then $\lambda \in \mathbf{C}_0$, $u_0 \in \boldsymbol{\mathcal{E}}(\lambda)$, and $y_0 = S(\lambda)u_0 \in \boldsymbol{\mathcal{E}}_*(\lambda)$, where

$$\begin{aligned}
S(\lambda) = D + C\,&\left((\xi_1\lambda_1 + \xi_2\lambda_2)I - (\xi_1A_1 + \xi_2A_2)\right)^{-1}\\
&\widetilde{B}\xi_1\sigma_1 + \xi_2\sigma_2)\,\big|_{\boldsymbol{\mathcal{E}}(\lambda)} : \boldsymbol{\mathcal{E}}(\lambda) \to \boldsymbol{\mathcal{E}}_*(\lambda)
\end{aligned}$$

is the joint transfer function of the vessel (independent of the choice of $\xi_1, \xi_2 \in \mathbb{C}$ as long as the rsolvent exists).

Allowing for zero multiplicities, we may assume that the input and the output discriminant polynomials share the same irreducible factors with possibly different multiplicities r_i and r_{i*}. Under additional assumptions (maximality of determinantal representations), $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{E}}_*$ lift to vector bundles of ranks r_i and r_{i*} respectively on the corresponding desingularizing Riemann surfaces. If we assume now that the system is finite-dimensional, i.e., $\dim \mathcal{X} < \infty$, then it follows that the joint transfer functions lifts, for each irredicible component of the discriminant curve, to a meromorphic budle map between kernel bundles (if the vessel is minimal $\Longleftrightarrow$ controllable and observable then the poles of the joint transfer function coincide exactly with the joint spectrum of A_1 and A_2).

The theory of determinantal representations identifies exactly, up to isomorphism, kernel bundles of determinantal representations: up to a certain twist they are isomorphic to vector bundles of the form $\mathbf{V}_\chi \otimes \Delta$, where $\mathbf{V}_\chi$ is a flat vector bundle corresponding to some representation χ of the fundamental group and Δ is a line bundle of differentials of order 1/2 (a square root of the canonical bundle), so that $h^0(\mathbf{V}_\chi \otimes \Delta) = 0$. The corresponding collection, one for veach irreducible component of the discriminant curve, of meromorphic bundle maps $\mathbf{V}_\chi \otimes \Delta \to \mathbf{V}_{\chi_*} \otimes \Delta$, is called the normalized joint transfer function of the vessel.

[1] So the discrete-time version yields a tool for spectral analysis of pairs of commuting nonunitary operators (primarily, commuting contractions).

THE REALIZATION PROBLEM

Assume for simplicity that we have only one irreducible component. The realization problem is then as follows: we are given a compact Riemann surface X together with a pair of meromorphic functions y_1 and y_2 on X that generate the whole field of meromorphic functions and determine therefore a birational embedding of X into $\mathbb{P}^2$ as an irreducible projective curve $\mathbf{C}$. We are also given a pair of vector bundles $\mathbf{V}_\chi \otimes \Delta$ and $\mathbf{V}_{\chi_*} \otimes \Delta$ on X as above, and a meromorphic bundle map $T\colon \mathbf{V}_\chi \otimes \Delta \to \mathbf{V}_{\chi_*} \otimes \Delta$ which is holomorphic at the poles of y_1 and y_2. We want to construct a vessel with discriminant curve $\mathbf{C}$, input and output determinantal representations with kernel bundles isomorphic to $\mathbf{V}_\chi \otimes \Delta$ and $\mathbf{V}_{\chi_*} \otimes \Delta$ respectively, and normalized joint transfer function T.

One approach to this problem, worked out essentially in Livšic–Kravitsky–Markus–Vinnikov (1995) in the infinite dimensional conservative setting, proceeds by first constructing the two determinantal representations and using the so called restoration formula to recover from T the characteristic function of a 1D colligation. One then applies the usual realization theorem. Another approach, worked out in Ball and Vinnikov (1996) in case T has only simple poles, is the analogue of the so called Gilbert realization — it constructs the vessel explicitly by factoring the residues.

Our approach will be very different: we will construct both the state space and the input/output spaces of the vessel as spaces of meromorphic sections of the bundle $\mathbf{V}_{\chi_*} \otimes \Delta$ using the meromorphic bundle map T and the functions y_1, y_2, analogously to the usual construction of the Hankel realization.

REFERENCES

J.A. Ball and V. Vinnikov, Zero-pole interpolation for matrix meromorphic functions on an algebraic curve and transfer functions of 2D systems, *Acta Appl. Math.* 45, 239–316 (1996).

J. A. Ball and V. Vinnikov, Overdetermined Multidimensional Systems: State Space and Frequency Domain Methods, *Mathematical Systems Theory* (D. Gilliam and J. Rosenthal, eds.), Inst. Math. and its Appl. Volume Series, Vol. 134, Springer-Verlag, New York (2003), 63–120.

M. S. Livšic, Commuting nonselfadjoint operators and solutions of systems of partial differential equations generated by them, *Soobshch. Akad. Nauk Gruzin. SSR* **91** (1978), no. 2, 281–284, **MR** 80m:47008. In Russian.

M. S. Livšic, Operator waves in Hilbert space and related partial differential equations, *Int. Eq. Oper. Th.* **2** (1979), no. 1, 25–47.

M. S. Livšic, The inverse problem for the characteristic functions of several commuting operators, *Int. Eq. Oper. Th.* **2** (1979), no. 2, 264–286.

M. S. Livšic, N. Kravitsky, A. S. Markus, and V. Vinnikov, *Theory of commuting nonselfadjoint operators*, Mathematics and Its Applications, vol. 332, Kluwer, Dordrecht, 1995.

M. S. Livšic, *Commuting nonselfadjoint operators and a unified theory of waves and corpuscles*, *Operator Theory: Adv. Appl.* 98, pp. 163–185 (1997).

M. S. Livšic, *Vortices of 2D systems*, *Operator Theory: Adv. Appl.* 123, 7–41 (2001).

M. S. Livšic, *Chains of space-time open systems and DNA*, *Operator Theory: Adv. Appl.* 134, 319–336 (2002).

V. Vinnikov, *Commuting operators and function theory on a Riemann surface*, Holomorphic Spaces and Their Operators (S. Axler, J. McCarthy, and D. Sarason, eds.), Math. Sci. Res. Inst. Publ., vol. 33, Cambridge University Press, Cambridge, 1998, pp. 445–476.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A tale of two cones: Psd vs Sos in equivariant situations

Charu Goel * **Salma Kuhlmann** **

* *Indian Institute of Information Technology Nagpur, India (e-mail: charugoel@iiitn.ac.in).*
** *University of Konstanz, Germany (e-mail: salma.kuhlmann@uni-konstanz.de)*

Abstract: The relationship between the cone of positive semidefinite (psd) real forms and its subcone of sums of squares (sos) of forms is of fundamental importance in real algebraic geometry and optimization, and has been studied extensively (see for instance Marshall (2008)). The study of this relationship goes back to the 1888 seminal paper of Hilbert, where he gave a complete characterisation of the pairs $(n, 2d)$ for which a psd n-ary $2d$-ic form can be written as sos. In this talk we discuss how this relationship changes under the additional assumptions of invariance on the given forms, i.e. when we consider the induced action of a real finite reflection group on the ring of polynomials. We will see that in equivariant situations Hilbert's classification does not remain true in general and depends on the group action, the degree and the number of variables.

Keywords: Positive semidefinite forms, Sums of squares, Symmetric forms, Test sets, Extremal rays

1. INTRODUCTION

Hilbert (1888) studied the inclusion $\mathcal{P}_{n,2d} \supseteq \Sigma_{n,2d}$, where $\mathcal{P}_{n,2d}$ and $\Sigma_{n,2d}$ are respectively the cones of psd and sos forms of degree $2d$ in n variables. He proved that:

$$\mathcal{P}_{n,2d} = \Sigma_{n,2d} \text{ if and only if } n = 2, d = 1, \text{ or } (n, 2d) = (3, 4).$$

In order to establish that $\Sigma_{n,2d} \subsetneq \mathcal{P}_{n,2d}$, he demonstrated that $\Sigma_{3,6} \subsetneq \mathcal{P}_{3,6}$, $\Sigma_{4,4} \subsetneq \mathcal{P}_{4,4}$, thus reducing the problem to these two basic cases using an argument to increase the number of variables and degree of a given psd not sos form while simultaneously preserving the psd not sos property. In these two cases Hilbert described a method to produce examples of psd not sos forms, which was "elaborate and unpractical" (see Choi et al (1977)), so no explicit examples appeared in literature for next 80 years. The first explicit examples of psd not sos forms in these two cases were found by Motzkin (1967) and Robinson (1969), in the late 1970's. Subsequently more examples were given by Choi-Lam (see Choi (1975), Choi et al (1976), Choi et al (1977)), Reznick (see Reznick (1989)) and Schmüdgen (see Schmüedgen (1979)).

In 1976, Choi and Lam (see Choi et al. (1977)) considered the same inclusion for *symmetric forms*, i.e forms invariant under the action of the symmetric group S_n. As an analogue of Hilbert's approach, they demonstrated that establishing the strict inclusion for all $n \geq 3, 2d \geq 4$ and $(n, 2d) \neq (3, 4)$ reduces to show it just for the pairs $(n, 2d) = (3, 6), (n, 4)_{n\geq 4}$, by using a trick that increases the degree – however not the number of variables – of a given psd not sos symmetric form by simultaneously preserving the psd not sos symmetric property. Assuming the existence of psd not sos symmetric n-ary quartics for $n \geq 5$, they showed that Hilbert's characterisation above remains unchanged. Recently, we (see Goel et al (2016)) constructed explicitly these quartic forms, thus completing their proof. For this we used test set for (positivity of) symmetric quartics, that was originally given by Choi-Lam-Reznick (see Choi et al (1980)) and later generalized by Timofte (see Timofte (2003)) for symmetric polynomials of degree $2d$ in n variables.

Recently, we studied systematically the above inclusion of cones for *even symmetric forms*, i.e. forms invariant under the action of the group $S_n \times \mathbb{Z}_2^n$. The idea was to develop an analogue of reduction to basic cases, in the same spirit as Hilbert and Choi-Lam. Choi-Lam-Reznick (see Choi et al (1987)) and Harris (see Harris (1999)) established that Hilbert's characterisation is no longer true for even symmetric forms; indeed equality of these cones holds also for the pairs $(n, 4)_{n\geq 4}$ and $(3, 8)$. Moreover, they gave psd not sos even symmetric examples for the pairs $(n, 6)_{n\geq 3}, (3, 10)$ and $(4, 8)$. Building up on their work, we (see Goel et al (2017)) established strict inclusion for the pairs $(3, 2d)_{d\geq 6}, (n, 8)_{n\geq 5}, (n, 2d)_{n\geq 4, d\geq 5}$, and proved that it suffices for all the remaining cases [i.e. for all $n \geq 3, 2d \geq 6$ and $(n, 2d) \neq (3, 8)$]. For this we introduced as our leading tool a "Degree Jumping Principle"(that increases the degree of a given psd not sos even symmetric form while simultaneously preserving the psd not sos even symmetric property) and constructed explicit counterexamples for the pairs $(n, 8)_{n\geq 5}, (n, 10)_{n\geq 4}, (n, 12)_{n\geq 4}$. This let us to a complete resolution of all remaining open cases, thus providing a complete analogue of Hilbert's theorem for even symmetric forms (see Goel et al (2017)), namely,

"*an even symmetric n-ary 2d-ic psd form is sos if and only if* $n = 2$ *or* $d = 1$ *or* $(n, 2d) = (n, 4)_{n\geq 3}$ *or* $(n, 2d) = (3, 8)$".

Recently, Blekherman-Riener (2020) studied the asymptotic behaviour of symmetric psd forms and symmetric sos forms when the degree $2d$ is fixed and the number of variables n grows. For degree 4 they showed that the difference between symmetric psd forms and sos forms asymptotically goes to zero. For $n \geq 4$, they gave *symmetric n-ary quartic forms which lies on the boundary of the symmetric sos cone but not on the boundary of the symmetric psd cone*. They used representation theory of the symmetric group and examined possible kernels of an extreme ray of the dual cone of symmetric n-ary sos quartics which does not come from a point evaluation. They developed this method for the symmetric situation from the work in Blekherman (2012) for the non-symmetric situation [where for $(3, 6)$ and $(4, 4)$ cases, the author constructed quadratic forms that span extremal rays of the dual cone of $\Sigma_{n,2d}$ but are not point evaluations]. Recently, Debus (2019) used methods similar to the ones in Blekherman et al. (2020), and *proved equality of psd and sos quaternary quartics invariant under the diagonal action of $D_{2.4} \times D_{2.4}$*, where the Dihedral group $D_{2.m}$ is the symmetry group of a regular m-gon.

2. PSD VS SOS IN EQUIVARIANT SITUATIONS

The problem of finding sos decompositions of a polynomial invariant under the action of a finite group was studied by Gatermann-Parrilo (2003) and by Cimpric et al. (2009) (for reductive groups). Further, it is interesting as well as important to find the explicit description of the cone of invariant sos forms; this is done for invariance under a finite group generated by pseudo reflections and octahedral group by Vallentin et al. (2017).

As discussed in section 1 above, taking invariance under a bigger group results in equality of the cones of invariant psd and invariant sos forms for more number of pairs. This naturally opens up an idea to *investigate a wider generalization of analogues of Hilbert's theorem for forms invariant under other group actions.*

One possible approach tries to *develop more sophisticated arguments and tools for the invariant forms under consideration*, along the same lines as Hilbert (1888), Choi-Lam (1977) and Goel-Kuhlmann-Reznick (2017), in particular

- reduction to basic cases
- indefinite irreducible multipliers invariant under the considered group G (like Degree jumping principle)
- counterexamples in basic cases
- generalization of Timofte's degree principle (see Timofte (2003)) to invariant forms

This is very important since all these tools (in particular test sets for positivity of symmetric polynomials) played an important role in establishing the analogues of Hilbert's theorem for symmetric and even symmetric forms. An adaptation of Riener's algebraic proof (see Riener (2012)) of Timofte's theorem would be necessary, depending on the invariants of the group action. In this spirit we are investigating the inclusion of cones for forms invariant under the action of finite reflection groups and Lie groups, using a recent generalization of Timofte's degree principle for these groups given by Acevedo-Velasco (2015) and Friedl-Riener-Sanyal (see Friedl et al. (2016)).

Another possible approach is to see how the representation theory of these groups allows to use the symmetry inherent in these cones to give more efficient descriptions, in particular using the methods similar to the ones in Blekherman-Riener (2012) and Debus (2019).

Any real reflection group can be identified with a direct product of essential reflection groups G. According to Coxeter classification (see Humphreys (1992)), these G includes four infinite families of irreducible reflection groups $A_{n-1}, B_n, D_n, I_2(m)$, and the six exceptional reflection groups $E_6, E_7, E_8, F_4, H_3, H_4$. In a joint work with Sebastian Debus and Cordian Riener, we are investigating inclusion of cones for forms invariant under the action of finite reflection groups, using the second approach.

For the first approach we needed "test sets" a la Timofte. On the other hand with the second approach that we suggested above, it would circumvent the need of a currently non existing general Timofte result for reductive groups (although such a general result would be interesting of course in its own right).

In this talk we will focus especially on the results for forms invariant under the action of A_{n-1}, B_n and D_n (based on joint work with Sebastian Debus and Cordian Riener). The case of forms invariant under the action of $I_2(m)$ is trivial, since the $I_2(m)$-invariant forms are bivariate and thus by Hilbert's characterisation these forms are psd if and only if they are sos.

3. CONCLUSION

It is noteworthy that the first counterexamples (i.e. psd not sos ternary sextics and quaternary quartics) substantiating Hilbert's 1888 theorem were given almost 80 years later in 1967. Moreover, the results on equality and strict inclusions of the cones of psd and sos forms (respectively symmetric, even symmetric forms) by Hilbert (respectively Choi-Lam-Reznick, Harris and Goel-Kuhlmann-Reznick) were building stones in establishing Hilbert's theorem (respectively its analogue for symmetric and even symmetric forms). Thus, given a finite group G, establishing equality or strict inclusion of cones of invariant psd and invariant sos forms for any (n, 2d) will be a novel contribution in this research area and would have a strong impact on the applications of sums of squares.

REFERENCES

J. Acevedo and M. Velasco. Test sets for nonnegativity of polynomials invariant under a finite reflection group. Journal of Pure and Applied Algebra, 220 (8), (2016), 2936-2947.

G. Blekherman. Nonnegative polynomials and sums of squares. Journal of the American Mathematical Society, 25(3), (2012), 617–635.

G. Blekherman and C. Riener. Symmetric nonnegative forms and sums of squares. Discrete and Computational Geometry, Volume 65 (3), (2020), 764-799.

J. Cimpric, S. Kuhlmann, C. Scheiderer. Sums of squares and moment problems in equivariant situations. Trans. Am. Math. Soc. 361 (2009), 735-765.

M. D. Choi. Positive semidefinite biquadratic forms. Linear Algebra Appl. 12 (1975), 95-100.

M.D. Choi and T.Y. Lam. An old question of Hilbert. Proc. Conf. quadratic forms, Kingston 1976, Queen's Pap. Pure Appl. Math. 46, (1977), 385-405.

M.D. Choi and T.Y. Lam. Extremal positive semidefinite forms. Math. Ann. 231, no.1 (1977), 1-18.

M.D. Choi, T.Y. Lam, B. Reznick. Symmetric quartic forms. Unpublished, 1980.

M.D. Choi, T.Y. Lam, B. Reznick. Even symmetric sextics. Math. Z. 195 (1987), 559-580.

S. Debus. Non-negativity versus Sums of squares in equivariant situations. Master thesis, University of Vienna, 2019.

M. Dostert, C. Guzmán, F. M. de Oliveira Filho, F. Vallentin. New Upper Bounds for the Density of Translative Packings of Three-Dimensional Convex Bodies with Tetrahedral Symmetry. Discrete Comput. Geom. 58 (2017), 449-481.

T. Friedl, C. Riener, R. Sanyal. Reflection groups, arrangements, and invariant real varieties. Proceedings of the AMS 146 (3) (2018), 1031–1045.

K. Gatermann and P. A. Parrilo. Symmetry groups, semidefinite programs, and sums of squares. J. Pure Appl. Algebra, 192, (1–3) (2004), 95–128.

C. Goel, S. Kuhlmann, B. Reznick. On the Choi-Lam analogue of Hilbert's 1888 theorem for Symmetric forms. Linear Algebra and its Applications, 496 (2016), 114-120.

C. Goel, S. Kuhlmann, B. Reznick. The analogue of Hilbert's 1888 theorem for Even Symmetric Forms. Journal of Pure and Applied Algebra, 221 (2017), 1438-1448.

W. R. Harris. Real even symmetric ternary forms. J. Algebra 222, no. 1 (1999), 204-245.

D. Hilbert. Ueber die Darstellung definiter Formen als Summe von Formenquadraten. Math. Ann., 32 (1888), 342-350; Ges. Abh. 2, 154-161, Springer, Berlin, reprinted by Chelsea, New York, 1981.

J. E. Humphreys. Reflection groups and Coxeter groups. Cambridge University Press, 1992.

M. Marshall Positive Polynomials and Sum of Squares. Vol. 146, Mathematical Surveys and Monographs, AMS, 2008.

T. S. Motzkin. The arithmetic-geometric inequality. in Inequalities, Oved Shisha (ed.) Academic Press (1967), 205-224.

B. Reznick. Forms derived from the arithmetic-geometric inequality. Math. Ann. 283 (1989), 431-464.

C. Riener. On the degree and half-degree principle for symmetric polynomials. Journal of Pure and Applied Algebra, vol. 216, no. 4 (2012), 850-856.

R. M. Robinson. Some definite polynomials which are not sums of squares of real polynomials. Selected questions of algebra and logic, Acad. Sci. USSR (1973), 264-282, Abstracts in Notices Amer. Math. Soc. 16 (1969), p. 554.

K. Schmüdgen. An example of a positive polynomial which is not a sum of squares of polynomials. A positive, but not strongly positive functional. Math. Nachr. 88 (1979), 385-390.

V. Timofte. On the positivity of symmetric polynomial functions. Part I: General results. J. Math Anal. Appl. 284 (2003), 174-190.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Existence of best low rank approximations for positive definite tensors

Eric Evert. * **Lieven De Lathauwer.** **

* *Group Science, Engineering and Technology
KU Leuven Kulak, E. Sabbelaan 53, 8500 Kortrijk, Belgium,
(e-mail: eric.evert@kuleuven.be)*
** *Group Science, Engineering and Technology
KU Leuven Kulak, E. Sabbelaan 53, 8500 Kortrijk, Belgium,
(e-mail: lieven.delathauwer@kuleuven.be)*

Abstract: The best low rank tensor approximation problem occurs in a wide variety of applications; however, this problem is strictly speaking not well posed. Indeed, best low rank tensor approximations can fail to exist. In the case that a best low rank approximation fails to exist, computing a near optimal low rank approximation is highly numerically ill-conditioned. In this talk we will consider the best low rank approximation problem for the special class of tensors which are positive definite. We will show that the set of low rank tensors that are positive definite is relatively closed as a subset of the set of tensors that are positive definite. Using this fact, we will provide a deterministic bound for the existence of a best low rank approximation of a positive definite tensor. We will illustrate through numerical experiments that our bound is highly predictive of numerical errors when attempting to compute a best low rank approximation of a measured tensor.

Keywords: Tensors, Canonical polyadic decomposition, Best low rank approximation, Positive definite

1. INTRODUCTION

Tensors, or multiindexed arrays, play an important role in fields such as machine learning and signal processing. These higher-order generalizations of matrices allow for preservation of higher-order structure present in data, and low rank decompositions of tensors allow for recovery of underlying information. One of the most popular decompositions for tensors is the canonical polyadic decomposition (CPD) which expresses a tensor as a sum of rank one tensors.

Tensor decompositions are widely used applications as they many desirable qualities which are distinct from matrix decompositions. For example, a decomposition of a low rank tensor is generically unique. This essential uniqueness allows for extraction of component information from a measured tensor of interest and makes tensor decomposition a valuable tool in applications such as blind source separation.

A common setting in practice is to have access to a noisy measurement of some low rank signal tensor of interest. This measurement itself does not have low rank, so one must compute a best low rank CPD approximation of the measured tensor. However, this approximation problem is ill-posed as the set of tensors of rank less than or equal to R is in general not closed when $R > 1$.

In the case a tensor does not have a best rank R approximation, then as one's rank R approximation improves, the norm of rank-1 tensors in the rank R approximation must approach infinity. This phenomena known as "diverging components". These diverging components present a serious numerical issue when one wishes to obtain component information from a tensor which does not have a best low rank approximation.

One particularly important setting is where the tensor of interest is positive (semi)definite. For example, positive definite tensors in the form of higher-order statistics play a key role in blind source separation. In addition, there are many applications in which a tensor models a nonlinear function which has positive evaluation on any input. The tensor used to model such a function will be positive definite.

In this talk, we show that the set of "low rank" positive definite tensors is relatively closed as a subset of the set of positive definite tensors. Using this result, we produce present a deterministic guarantee for the existence of a best low rank approximation in the form of a spectral norm bound on measurement error. Furthermore we show that this bound is computable via semidefinite programming and sharp in certain settings.

⋆ This research received funding from (1) Flemish Government: This work was supported by the Fonds de la Recherche Scientifique–FNRS and the Fonds Wetenschappelijk Onderzoek–Vlaanderen under EOS Project no 30468160 (SeLMA) and under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme; (2) KU Leuven Internal Funds C16/15/059; (3) NSF grant DMS-1500835

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

2. POSITIVE DEFINITE TENSORS

In this talk, a tensor is a multiinedexed array $\mathcal{T}$ of size $R \times \ldots \times R$. Here the integer R occurs a total of N times where N is even. We denote this space by $(\mathbb{R}^R)^{\otimes N}$. We say a tensor $\mathcal{T} \in (\mathbb{R}^R)^{\otimes N}$ is symmetric if $\mathcal{T}$ if the entries of $\mathcal{T}$ are invariant under a permutation of indices. That is, $\mathcal{T} \in (\mathbb{R}^R)^{\otimes N}$ is symmetric if for all permutations π acting on the set $\{1, \ldots, N\}$, one has

$$\mathcal{T}(i_1, i_2, \ldots, i_N) = \mathcal{T}(i_{\pi(1)}, i_{\pi(2)}, \ldots, i_{\pi(N)}).$$

for all $i_1, i_2, \ldots, i_N$.

Given a symmetric tensor $\mathcal{T} \in (\mathbb{R}^R)^{\otimes N}$, a common goal is to compute the (symmetric) canonical polyadic decomposition (CPD) of $\mathcal{T}$. That is, one wishes to decompose

$$\mathcal{T} = \sum_{r=1}^{R} \delta_r (\mathbf{u}_r \otimes \ldots \otimes \mathbf{u}_r) \tag{1}$$

where R is as small as possible. Here $\mathbf{u}_r \in \mathbb{R}^R$ and $\delta_r = \pm 1$ for each $r = 1, \ldots, R$. The product $\otimes$ denotes the outer product of vectors. In particular, the tensor $\mathbf{u}_r \otimes \ldots \otimes \mathbf{u}_r$ has $i_1, i_2, \ldots, i_{N+1}$ entry equal to

$$\mathbf{u}_r(i_1) \ldots \mathbf{u}_r(i_N),$$

where $\mathbf{u}_r(i_n)$ denotes the i_nth entry of $\mathbf{u}_r$. When R is as small as possible in equation (1), we say the tensor $\mathcal{T}$ has rank R. From here on, will use $(\mathbf{u_r})^{\otimes N}$ to denote the symmetric rank one tensor $\mathbf{u}_r \otimes \ldots \otimes \mathbf{u}_r$.

We let $\langle \cdot, \cdot \rangle$ denote the usual Frobenius inner product on $(\mathbb{R}^R)^{\otimes N}$. That is, given tensors $\mathcal{T}, \mathcal{S} \in (\mathbb{R}^R)^{\otimes N}$, one has

$$\langle \mathcal{T}, \mathcal{S} \rangle = \sum \mathcal{T}(i_1, i_2, \ldots, i_N) \mathcal{S}(i_1, i_2, \ldots, i_N).$$

Define quantities $\lambda_{\min}(\mathcal{T})$ and $\lambda_{\max}$ by

$$\lambda_{\min}(\mathcal{T}) = \min_{\|\mathbf{u}\|_2 = 1} \langle \mathcal{T}, (\mathbf{u})^{\otimes N} \rangle \qquad \lambda_{\max}(\mathcal{T}) = \max_{\|\mathbf{u}\|_2 = 1} \langle \mathcal{T}, (\mathbf{u})^{\otimes N} \rangle.$$

We say a tensor $\mathcal{T}$ is positive definite if $\mathcal{T}$ is symmetric and if $\lambda_{\min}(\mathcal{T}) > 0$.

The first main result of the talk will be to show that the set of rank R positive definite tensors in $(\mathbb{R}^R)^{\otimes N}$ is relatively closed as a subset of the set of positive definite tensors in $(\mathbb{R}^R)^{\otimes N}$. Intuitively this is accomplished by showing that a rank R positive definite tensor in $(\mathbb{R}^R)^{\otimes N}$ must be expressed as a positive coefficient sum of positive semidefinite rank one tensors. This prevents the occurrence of diverging components for tensors in a neighborhood of $\mathcal{T}$ which in turn leads to the relative closedness of the set of rank R positive definite tensors.

We use this result to provide a deterministic guarantee for the existence of a best low rank approximation of a perturbation of a rank R positive definite tensor $\mathcal{T}$. In particular, we show that if $\mathcal{E} \in (\mathbb{R}^R)^{\otimes N}$ is a symmetric tensor which satisfies

$$\lambda_{\max}(\mathcal{E})/2 < \lambda_{\min}(\mathcal{T}),$$

then $\mathcal{T} + \mathcal{E}$ has a best rank R approximation among the set of symmetric tensors.

For tensors arising in applications where the tensor of interest is expected to be positive definite, this existence result can be interpreted as saying that the measured tensor $\mathcal{T} + \mathcal{E}$ has a best low rank approximation so long as the conditioning of $\mathcal{T}$ and the signal to noise ratio of $\mathcal{T}$ to $\mathcal{E}$ is good enough so that near optimal low rank approximations to $\mathcal{T} + \mathcal{E}$ are positive definite. Thus, a best low rank approximation exists so long as it exhibits the properties that are expected for the setting.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A Mayer Form for Finite Horizon Hybrid Optimal Control Problems

Ricardo G. Sanfelice and Berk Altin

***Abstract*— We consider finite horizon optimal control problems for hybrid plants that are modeled as hybrid equations. To determine key properties of the problem, such as existence and regularity of the optimal cost, we formulate a Mayer form that is tailored to hybrid systems. Within the setting of nominally outer well-posed hybrid plants, and under mild (and standard) regularity conditions, establishing existence of optimal solutions and nice (upper semicontinuous and continuous) dependence of the optimal cost is enabled by the proposed Mayer form. The advantage of the proposed approach is that it does not require additional properties that are typically required in the literature, such as assumptions on the continuous dynamics or that the terminal cost is a control Lyapunov function on the terminal constraint set. The proposed new form is illustrated in examples.**

I. Introduction

Models and algorithms characterized by the interplay of continuous-time dynamics and instantaneous changes have become prevalent due to their capabilities of leading to solutions to control problems that classical techniques cannot solve, or simply do not apply. These advances have been enabled by the modeling, analysis, and design techniques for *hybrid dynamical systems*. A hybrid dynamical system, or just a *hybrid system*, is a dynamical system that exhibits characteristics of both continuous-time and discrete-time dynamical systems.

Numerous tools are available in the literature for the study of hybrid systems [1]–[6]. The literature is rich in tools for the analysis of reachability [7]–[9], asymptotic stability [1], [3], [5], forward invariance [10], [11], and control design [6]. On the other hand, optimality for hybrid systems is much less mature.

Initial results on optimality of trajectories over finite horizons were developed in [12], including a maximum principle for optimality, for a class of switched systems. This result was extended in [13], [14] to a broader class of systems, one allowing for state resets – the models considered are in the spirit of hybrid automata. More recently, linear-quadratic control for a class of hybrid systems with a sample-and-hold structure was considered in [15], [16]. In particular, the development in [15] is within the hybrid inclusions framework of [5], [6], for the special case when the continuous dynamics are modeled by a differential equation that is linear and the discrete dynamics are governed by a linear difference equation. The problem of guaranteeing existence of optimal control inputs for a class of hybrid systems was studied in [17]. The hybrid inclusions framework is employed in [17] and the conditions for existence of optimal control inputs require the continuous dynamics of the system to be governed by a differential equation whose right-hand side is affine in the control input. Optimality of static state-feedback laws for hybrid inclusions with continuous and discrete dynamics modeled by (single-valued) nonlinear maps was studied in [18]. Infinitesimal conditions involving a Lyapunov-like function are presented in [18] to guarantee optimality over the infinite (hybrid) horizon. The finite horizon optimization problem for the same broad class of hybrid systems was formulated and developed in a sequence of papers leading to a model predictive control framework; see [19]–[22].

Though the advances cited above have contributed to optimal control for hybrid systems, some of the key properties of the optimal control problem associated to general hybrid systems, wherein trajectories are constrained to evolve continuously (*flow*) in certain regions of the state space and to exhibit instantaneous changes (*jump*) under certain conditions, have not been yet revealed in the literature. Specifically, the regularity properties of the optimal cost, in particular, (semi) continuous dependence of the optimal cost and optimal trajectories on the constraints on where the trajectories can flow or jump have not yet been investigated. Very importantly, conditions enabling the approximations of the optimal cost in a continuous manner are not available in the literature. Indeed, results that permit relating the effect of varying parameters and initial conditions when they approach nominal values, the expectation being that the optimal cost also approaches its nominal value, are missing. Understanding such a dependency is critical due to the fact that it is unavoidable to numerically compute trajectories (hence the optimal trajectories) without error [23], [24].

II. Overview of the Presentation

We consider concrete finite horizon optimization problems for hybrid plants given by

$$\mathcal{H}_P \begin{cases} \dot{x}_P \in F_P(x_P, u) & (x_P, u) \in C_P \\ x_P^+ \in G_P(x_P, u) & (x_P, u) \in D_P \end{cases} \tag{1}$$

where C_P is the flow set, F_P is the flow map, D_P is the jump set, and G_P is the jump map. A solution of $\mathcal{H}_P$ is defined by a pair (called a *solution pair*) $(t, j) \mapsto (x_P(t, j), u(t, j))$ on a hybrid time domain $\mathrm{dom}(x_P, u)$ satisfying the dynamics of $\mathcal{H}_P$, in a similar manner as the way a solution of the (closed-loop) hybrid system $\mathcal{H}$ is defined in [5]. Given a solution pair (x_P, u) with compact domain, the associated

R. G. Sanfelice is with the Department of Electrical and Computer Engineering, University of California, Santa Cruz, CA 95064. Email: ricardo@ucsc.edu. Research partially supported by NSF Grants no. ECS-1710621, CNS-2039054, and CNS-2111688, by AFOSR Grants no. FA9550-19-1-0053, FA9550-19-1-0169, and FA9550-20-1-0238, and by ARO Grant no. W911NF-20-1-0253.

cost is defined by

$$\left(\sum_{j=0}^{J}\int_{t_j}^{t_{j+1}} L_{C_P}(x_P(t,j),u(t,j))\,dt\right)+\left(\sum_{j=0}^{J-1} L_{D_P}(x_P(t_{j+1},j),u(t_{j+1},j))\right)+V(x_P(T,J)) \tag{2}$$

where t_j is the j-th jump time and $(T,J)\in \mathrm{dom}(x_P,u)$ is the terminal time, i.e.,

$$\mathrm{dom}(x_P,u)=\bigcup_{j=0}^{J}([t_j,t_{j+1}]\times\{j\})$$

and $T=T_{J+1}$. In (2), the first term L_{C_P} is the stage cost capturing the cost over intervals of flows, L_{D_P} is the stage cost capturing the cost to jump, and V is the terminal cost.

The constructions presented above lead to the following finite horizon hybrid optimization problem.

Problem: Given a hybrid system $\mathcal{H}_P$ as in (1), a stage cost for flows L_{C_P}, a stage cost for jumps L_{D_P}, a terminal cost V, a closed set X_P, a hybrid time $(T,J)\in\mathbb{R}_{\geq 0}\times\mathbb{N}:=[0,\infty)\times\{0,1,\ldots\}$, and an initial condition ξ, find a solution pair (x_P,u) minimizing (2) subject to

- The initial condition constraint $x_P(0,0)=\xi$.
- The terminal constraint $x_P(T,J)\in X_P$.

Note that the flow and jump sets of $\mathcal{H}_P$ impose constraints that the solution pair needs to satisfy during flows and jumps, respectively. In fact, for the solution pair to exist up to hybrid time (T,J) it has to belong to C_P and D_P: as [6, Definition 2.29] indicates, (x_P,u) is a solution of $\mathcal{H}_P$ if

- $(x_P(0,0),u(0,0))\in\overline{C_P}\cup D_P$;
- For each $j\in\mathbb{N}$,

  $$(x_P(t,j),u(t,j))\in C_P$$

  for all $t\in \mathrm{int}I^j$ and

  $$\frac{dx_P}{dt}(t,j)\in F_P(x_P(t,j),u(t,j))$$

  for almost all $t\in I^j$, where $I^j:=\{t:(t,j)\in\mathrm{dom}(x_P,u)\}$;
- For each $(t,j)\in\mathrm{dom}(x_P,u)$ such that $(t,j+1)\in\mathrm{dom}(x_P,u)$,

  $$(x_P(t,j),u(t,j))\in D_P$$

  and

  $$x_P(t,j+1)\in G_P(x_P(t,j),u(t,j))$$

In this presentation, we determine key properties of the problem, such as existence and regularity of the optimal cost, by formulating a Mayer form that is tailored to hybrid systems. Under mild – and standard – regularity conditions, we show that when the hybrid system is well-posed, existence of optimal solutions and nice (upper semicontinuous and continuous) dependence of the optimal cost can be established using the proposed Mayer form. The results will be illustrated in examples.

REFERENCES

[1] J. Lygeros, K.H. Johansson, S.N. Simić, J. Zhang, and S. S. Sastry. Dynamical properties of hybrid automata. *IEEE Transactions on Automatic Control*, 48(1):2–17, 2003.

[2] A. van der Schaft and H. Schumacher. *An Introduction to Hybrid Dynamical Systems*. Lecture Notes in Control and Information Sciences, Springer, 2000.

[3] W. M. Haddad, V. Chellaboina, and S. G. Nersesov. *Impulsive and Hybrid Dynamical Systems: Stability, Dissipativity, and Control*. Princeton University, 2006.

[4] M.S. Branicky, V. S. Borkar, and S. K. Mitter. A unified framework for hybrid control: Model and optimal control theory. *IEEE Transactions on Automatic Control*, 43(1):31–45, 1998.

[5] R. Goebel, R. G. Sanfelice, and A. R. Teel. *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*. Princeton University Press, New Jersey, 2012.

[6] R. G. Sanfelice. *Hybrid Feedback Control*. Princeton University Press, New Jersey, 2021.

[7] J. Lygeros, C. Tomlin, and S. S. Sastry. Controllers for reachability specifications for hybrid systems. *Automatica*, 35:349–370, 1999.

[8] B. Altın and R. G. Sanfelice. Semicontinuity properties of solutions and reachable sets of nominally well-posed hybrid dynamical systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 5755–5760, 2020.

[9] P. Collins. Semantics and computability of the evolution of hybrid systems. *SIAM Journal on Control and Optimization*, 49(2):890–925, 2011.

[10] J.-P. Aubin, J. Lygeros, M. Quincampoix, S. S. Sastry, and N. Seube. Impulse differential inclusions: a viability approach to hybrid systems. *IEEE Transactions on Automatic Control*, 47(1):2–20, 2002.

[11] J. Chai and R. G. Sanfelice. Forward invariance of sets for hybrid dynamical systems (Part I). *IEEE Transactions on Automatic Control*, 64:2426–2441, 06/2019 2019.

[12] H. J. Sussmann. A maximum principle for hybrid optimal control problems. In *Proc. 38th IEEE Conference on Decision and Control*, pages 425–430, 1999.

[13] M. S. Shaikh and P. E. Caines. On the hybrid optimal control problem: Theory and algorithms. *IEEE Transactions on Automatic Control*, 52:1587–1603, 2007.

[14] A. Pakniyat and P. E. Caines. On the hybrid minimum principle: The hamiltonian and adjoint boundary conditions. *IEEE Transactions on Automatic Control*, 66(3):1246–1253, 2020.

[15] C. Possieri and A. R. Teel. Lq optimal control for a class of hybrid systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 604–609. IEEE, 2016.

[16] A. Cristofaro, C. Possieri, and M. Sassano. Linear-quadratic optimal control for hybrid systems with state-driven jumps. In *2018 European Control Conference (ECC)*, pages 2499–2504. IEEE, 2018.

[17] R. Goebel. Existence of optimal controls on hybrid time domains. *Nonlinear Analysis: Hybrid Systems*, 31:153 – 165, 2019.

[18] F. Ferrante and R. G. Sanfelice. Certifying optimality in hybrid control systems via lyapunov-like conditions. In *11th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2019)*, pages 245–250, 2019.

[19] B. Altın, P. Ojaghi, and R. G. Sanfelice. A model predictive control framework for hybrid dynamical systems. *IFAC-PapersOnLine*, 51(20):128–133, 2018. 6th IFAC Conference on Nonlinear Model Predictive Control NMPC 2018.

[20] B. Altın and R. G. Sanfelice. Asymptotically stabilizing model predictive control for hybrid dynamical systems. In *2019 American Control Conference (ACC)*, pages 3630–3635, July 2019.

[21] P. Ojaghi, Berk Altın, and R. G. Sanfelice. A model predictive control framework for asymptotic stabilization of discretized hybrid dynamical systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2356–2361, 2019.

[22] B. Altın and R. G. Sanfelice. Model predictive control for hybrid dynamical systems: Sufficient conditions for asymptotic stability with persistent flows or jumps. In *2020 American Control Conference (ACC)*, pages 1791–1796, 2020.

[23] A. M. Stuart and A.R. Humphries. *Dynamical Systems and Numerical Analysis*. Cambridge University Press, 1996.

[24] R. G. Sanfelice and A. R. Teel. Dynamical properties of hybrid systems simulators. *Automatica*, 46(2):239–248, 2010.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Realization of matrix monotone and matrix convex functions ⋆

Ryan Tully-Doyle *

* *California Polytechnic State University, San Luis Obispo, CA USA (e-mail: rtullydo@calpoly.edu).*

Keywords: matrix monotone functions, matrix convex functions, matrix valued functions, realization theory

A fundamental result of Löwner gives that a real-valued function is matrix monotone if and only if it extends to an analytic function with a rather rigid structure. Löwner's student Kraus essentially showed that functions that are matrix convex are analytic functions with a similarly rigid form.

Recall that Hermitian matrices have a partial ordering where $A \leq B$ if $0 \leq B - A$ (that is, $B - A$ is positive semidefinite). Suppose that $f : (a, b) \to \mathbb{R}$. A function is called **matrix monotone** if $f(A) \leq f(B)$ whenever $A \leq B$ and the spectrum of A, B is in (a, b). The classical Löwner theorem gives the following connection between monotonicity and analyticity. Let $\mathbb{H}$ denote the complex upper half plane.

Theorem 1. (K. Löwner (1934)). A function $f(a, b) \to \mathbb{R}$ is matrix monotone if and only if f extends to a continuous function $f : \mathbb{H} \cup (a, b) \to \overline{\mathbb{H}}$ that is analytic on $\mathbb{H}$.

Functions that map $\mathbb{H}$ into itself are called **Pick functions**. Nevalinna showed that Pick functions are characterized by a straightforward integral representation in terms of a positive Borel measure μ:

Theorem 2. (Nevanlinna (1929)). Let $f : \mathbb{H} \to \mathbb{C}$. The function f is analytic and maps $\mathbb{H}$ to $\overline{\mathbb{H}}$ if and only if there exist $a \in \mathbb{R}$, $b \geq 0$ and a positive Borel measure μ on $\mathbb{R}$ where $\frac{1}{1+t^2}$ is μ-integrable such that

$$f(z) = a + bz + \int_{\mathbb{R}} \frac{1}{t - z} - \frac{t}{1 + t^2} \, d\mu(t)$$

for all $z \in \mathbb{H}$.

Now let $f : (a, b) \to \mathbb{R}$ be a function. We say that f is **matrix convex** if

$$f\left(\frac{A + B}{2}\right) \leq \frac{f(A) + f(B)}{2}$$

for all A, B self-adjoint with spectrum in (a, b). Matrix convex functions demonstrate the same essential rigidity as matrix monotone functions.

Theorem 3. (Kraus (1936)). Let $f : (-1, 1) \to \mathbb{R}$. f is matrix convex if and only if

$$f(x) = a + bx + \int_{[-1,1]} \frac{x^2}{1 + tx} \, d\mu(t)$$

where $a, b \in \mathbb{R}$ and μ is a finite measure supported on $[-1, 1]$. Note that all such functions analytically continue to the upper half plane - that is, matrix convex functions also continue as to a subset of Pick functions.

We are concerned with how to generalize these results to functions of several variables. A natural setting for the study of such functions turns out to be the rapidly developing area of noncommutative function theory, which concerns functions several noncommuting variables acting on domains of matrix tuples.

Define the **matrix universe** to be tuples of same sized matrices

$$M^d = \bigcup M_n(\mathbb{C})^d.$$

A **free set** $D \subseteq M^d$ satisfies

(1) $X, Y \in D \Rightarrow X \oplus Y \in D$
(2) X implies $U^*XU \in D$ whenever U is unitary.

A **(real) free function** $f : D \to M^1$ satisfies:

(1) $f(X \oplus Y) = f(X) \oplus f(Y)$
(2) $f(U^*XU) = U^*f(X)U$ whenever U is unitary.

In the context of functions of several noncommuting variables, we have a direct generalization of this single-variable connection between monotonicity/convexity, analytic extension, and structured representation. We first consider the case of matrix monotone functions (a selection of relevant work in commutative and noncommutative variables can be found in Agler et al. (2012); Pascoe (2019, 2018); Palfia (2020); Pascoe and Tully-Doyle (2017, 2022, 2021)).

We restrict our attention to the noncommutative case. As in the single-variable setting, a noncommutative function f is said to be matrix monotone if $f(A) \leq f(B)$ whenever $A \leq B$.

Theorem 4. (Pascoe and Tully-Doyle (2022)). Let f be a locally bounded matrix monotone function defined on a convex free set of self-adjoints containing 0. Then there exist a scalar a_0, a vector Q, projections P_i, and a bounded self-adjoint contraction A such that

$$f(X) = a_0 + Q^*(A - \sum P_i X_i^{-1})^{-1} Q. \tag{1}$$

As in the classical case, the form of this "Nevanlinna representation" gives analytic continuation to a matrix upper half plane - that is, matrix monotone functions in

⋆ Partially supported by National Science Foundation DMS Analysis Grant 2055098

the noncommutative setting are those that analytically continue as Pick functions.

The idea of the proof goes through power series. The assumptions on f allow the application of the "royal road" idea, which uses a complex analytic approach to show that matrix monotone functions must be real analytic. Then a matrix monotone function f has a power series $f(X) = \sum c_\alpha X^\alpha$. Taking the derivative of f gives

$$Df(X)[H] = \sum c_{\beta^* x_i \alpha} X^{\beta^*} H_i X^\alpha$$

The monotonicity of f leads to the positivity of the x_i-localizing matrices $C_i = [c_{\beta^* x_i \alpha}]_{\alpha,\beta}$, which in turn allow the construction of the ambient Hilbert spaces $\mathcal{H}_i$ from which the projections P_i and the operator A are derived in the representation (1) .

The royal road approach to realization runs through a argument involving an application of the classical theorem and presupposes the form of the realization formula to be established. Another line of argument uses a Hilbert space approach (for example, in Agler et al. (2012) in the case of several commuting variables) to derive the representations. For a thorough treatment of the Hilbert space methods, (so-called operator analysis) see Agler et al. (2020). Löwner's theorem has recently received attention even in the classical case (see Simon (2019)).

We now consider the analogue of Kraus's theorem on matrix convex functions. As in the classical case, say that a noncommutative function f is matrix convex if

$$f\left(\frac{A+B}{2}\right) \leq \frac{f(A)+f(B)}{2}.$$

Matrix convex functions arise as an important object of study in the body of work surrounding linear matrix inequalities. In Helton et al. (2006), Helton, McCullough, and Vinnikov constructed a Kraus-like "butterfly realization" for noncommutative rational matrix convex functions. A similar argument to the proof of Theorem 4 gives that the result holds for general functions.

Theorem 5. (Pascoe and Tully-Doyle (2022)). Let f be a locally bounded matrix convex function defined on a convex free set of self-adjoints containing 0. Then there exist self-adjoint T_i, a vector Q, a scalar a_0, and a linear function L such that

$$f(X) = a_0 + L(X) + (\sum Q_i X_i)^*(I - \sum T_i X_i)^{-1}(\sum Q_i X_i).$$

As in the case of monotone functions, this representation also has an analytic continuation result, but in this case to a subset of a matrix upper half plane.

The study of noncommutative realizations for structured families of functions has developed to include, for example, partially matrix convex functions (see Jury et al. (2021)) and plurisubharmonic functions (see Dym et al. (2020); Pascoe (2021)). A noncommutative function is **partially matrix convex** if it is convex in one class of variables - that is

$$f\left(A, \frac{X+Y}{2}\right) \leq \frac{f(A,X)+f(A,Y)}{2}.$$

Combining the complex analytic "royal road" viewpoint of Pascoe and Tully-Doyle (2022) with the study of partially matrix convex rational functions in Dym et al. (2020) leads to the following realization theorem.

Theorem 6. (Jury et al. (2021)). A noncommutative rational function r in two variables a and x is partially matrix convex in x on self-adjoints near 0 if and only if there exists a vector rational function $\ell(a,x)$ that is linear in x and a matrix rational function $w(a)$ such that

$$r(a,x) = \ell(a,x)^* \sqrt{w(a)}(I - \sum \sqrt{w(a)} T_i x_i \sqrt{w(a)})^{-1} \sqrt{w(a)} \ell(a,x) + f(a,x)$$

where $f(a,x)$ is affine linear in x and the T_i are matrices.

Ultimately, the perspective underlying this line of work leads to a striking monodromy result (Pascoe (2021)) in the noncommutative setting.

REFERENCES

Agler, J., McCarthy, J., and Young, N.J. (2020). *Operator Analysis.* Cambridge University Press.

Agler, J., McCarthy, J., and Young, N. (2012). Operator monotone functions and Löwner functions of several variables. *Ann. of Math.*, 176, 1783–1826.

Dym, H., Helton, J.W., Klep, I., McCullough, S., and Volcic, J. (2020). Plurisubharmonic noncommutative rational functions. *J. Math. Anal. Appl.*

Helton, J.W., McCullough, S.A., and Vinnikov, V. (2006). Noncommutative convexity arises from linear matrix inequalities. *Journal of Functional Analysis*, 240(1), 105 – 191.

Jury, M., Klep, I., Mancuso, M., McCullough, S., and Pascoe, J.E. (2021). Noncommutative partially convex rational functions. *Rev. Mat. Iberoam.*

K. Löwner (1934). Über monotone Matrixfunktionen. *Math. Z.*, 38, 177–216.

Kraus, F. (1936). Über konvexe Matrixfunktionen. *Math. Z.*, 41, 18–42.

Nevanlinna, R. (1929). Über beschränkte Funktionen. *Ann. Acad. Sci. Fenn. Ser. A*, 32(7), 7–75.

Palfia, M. (2020). Löwner's theorem in several variables. *J. Math. Anal. Appl.*, 490(1).

Pascoe, J.E. (2018). The noncommutative Löwner theorem for matrix monotone functions over operator systems. *Linear Algebra and its Applications*, 541, 54 – 59.

Pascoe, J.E. (2019). Note on Löwner's theorem on matrix monotone functions in several commuting variables of Agler, McCarthy, and Young. *Monatsh. Math.*, 189, 377–381.

Pascoe, J.E. (2021). Noncommutative free universal monodromy, harmonic conjugates, and plurisubharmonicity. Preprint.

Pascoe, J.E. and Tully-Doyle, R. (2021). Automatic real analyticity and a regal proof of a commutative multivariate löwner theorem. *Proc. Amer. Math. Soc*, 149.

Pascoe, J.E. and Tully-Doyle, R. (2017). Free Pick functions: representations, asymptotic behavior and matrix monotonicity in several noncommuting variables. *J. Funct. Anal.*, 273(1), 283–328.

Pascoe, J.E. and Tully-Doyle, R. (2022). The royal road to automatic noncommutative real analyticity,

monotonicity, and convexity. To appear *Adv. Math., arXiV:1907.05875.*

Simon, B. (2019). Loewner's Theorem on Monotone Matrix Functions. *Springer.*

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Constrained Multiagent Dynamical systems and Hamilton-Jacobi-Bellman inequalities on the Wasserstein space

Marc Quincampoix *

* *Laboratoire de Mathématiques de Bretagne Atlantique, UMR CNRS 6205, Univ Brest 6 , avenue Victor Le Gorgeu, 29200 Brest, France. (e-mail: marc.quincampoix@univ-brest.fr)*

Abstract: Several optimal control problems in R^d, like systems with uncertainty, control of flock dynamics, or control of multiagent systems, can be naturally formulated in the space of probability measures in R^d. The compatibility of such control systems with a state constraint can be studied by an Hamilton-Jacobi-Bellman equation stated in the Wasserstein space of probability measure. We show that the dynamic is compatible with the constraint when the distance function satisfies the Hamilton Jacobi inequality in a suitable viscosity sense.

Keywords: Optimal control, optimal transport, Hamilton-Jacobi-Bellman equation, multi-agent.

1. INTRODUCTION

The studied multiagent system concerns a control systems with a so huge amount of agent that they are indistiguishable and only a statistical description of the position of the agents is available : at every time for every subset $A \subset R^d$, the fraction $\nu(A)$ of the total amount of agents that are present in A is known. So the state variable of the system is a Borel probabily measure on R^d (the set of such probability measure is denoted by $\mathcal{P}(R^d)$. Hence, the evolution of the controlled multi-agent system can be represented by the following two-scale dynamics

- *microscopic dynamics:* each agent's position at time t is given by $x(t)$, which evolves according to the dynamical system
$$\dot{x}(t) \in F(\mu_t, x(t)), \text{ for a.e. } t > 0\,, \tag{1}$$
where F is a set-valued map.Here each agent's dynamics is nonlocal since it depends also on the instantaneous configuration μ_t of the crowd of agents at time t, described by the probability measure $\mu_t \mathcal{P}(R^d)$.
- *macroscopic dynamics:* the evolution of the crowd of agents at time t is given by a time-depending measure $\mu_t \in \mathcal{P}(R^d)$ whose evolution satisfies the following *continuity equation*(cf Amb (2008))
$$\partial_t \mu_t + \text{div}(v_t \mu_t) = 0,\ t > 0, \tag{2}$$
coupled with the control constraint
$$v_t(x) \in F(\mu_t, x) \text{ for } \mu_t\text{-a.e. } x \in R^d \text{ for a.e. } t \geq 0. \tag{3}$$
which represents the possible velocity $v_t(x)$ for an agent at time t and at the position x.

Similar models have been studied in Ave (2018, 2021); Bard (2022); Bon (2021); Cav (2021); Jim (2020, 2021); Mar (2018).

⋆ The research of the third author has been partially supported by the Air Force Office of Scientific Research under award number FA9550-18-1-0254.

Here we investigate the existence of solution of the above multiagent system under a closed constraint $K \subset \mathcal{P}(R^d)$. This constraint could be satisfied in the both folool=wing meaning:

- The multiagent system is *viable* if and only if for any initial condition $\mu \in K$ *there exists* a solution $t \mapsto \mu_t$ of the controlled continuity equation (2)-(3) with $\mu_0 = \mu$ such that $\mu_t \in rK$ for all $t \geq 0$;
- The multiagent system is *invariant* if and only if for any μK and *for any* solution $t \mapsto \mu_t$ of the controlled continuity equation (2)-(3) with $\mu_0 = \mu$ we have $\mu_t \in K$ for all $t \geq 0$.

Following an idea firstly used in Buck (1998) in the framework of stochastic control systems we give an equivalent characterization of the viability and invariance in terms of of a suitable Hamilton Jacobi Bellman inequalities understood in the viscosity sense. This short note uses result obtained in collaboration with Cavagnari, Jimenez and Marigonda.

2. ASSUMPTIONS ON THE DYNAMICS

We denote by $\mathcal{P}_2(R^d)$ the subset of the elements $\mathcal{P}(R^d)$ with finite second moment, endowed with the 2-Wasserstein distance (cf e.g. Amb (2008)).

We make the following supposition on the set-valued map F

- (F_1) $F : \mathcal{P}_2(R^d) \times R^d \mapsto R^d$ is continuous with convex, compact and nonempty images, where on $\mathcal{P}_2(R^d) \times R^d$ we consider the metric $W_2(\mu_1, \mu_2) + |x_1 - x_2|$.
- (F_2) there exists $L > 0$, a compact metric space U and a continuous map $f : \mathcal{P}_2(R^d) \times R^d \times U \to R^d$ satisfying
$$|f(\mu_1, x_1, u) - f(\mu_2, x_2, u)| \leq L(W_2(\mu_1, \mu_2) + |x_1 - x_2|),$$

for all μ_1, μ_2 in $\mathcal{P}_2(R^d)$ and x_1, x_2 in R^d such that

$$F(\mu, x) = \{f(\mu, x, u) : u \in U\}.$$

Under these assumptions we know that the set of solution of (2)-(3) with $\mu_0 = \mu$ is nonempty and compact (cf Jim (2020, 2021); Mar (2018)).

Given a closed subset $K \subset \mathcal{P}_2(R^d)$, we define its distance function $d_K : \mathcal{P}_2(R^d) \mapsto R^+$ by

$$d_K(\mu) := \inf\{\, W_2(\mu, \nu),\ \ \nu \in K\,\}.$$

3. HAMILTON JACOBI INEQUALITIES

Under the above assumptions are now ready to state the main result

***Theorem** 3.1.* Let K be a closed subset of $\mathcal{P}_2(R^d)$

The constraint K is *viable* iff the function $\mu \mapsto d_K(\mu)$ is a viscosity supersolution of

$$(L+2)u(\mu) + H_F^{viab}(\mu, D_\mu u(\mu)) = 0,$$

where, for all $\mu \in \mathcal{P}_2(R^d)$, $p \in L^2_\mu(R^d; R^d)$,

$$H_F^{viab}(\mu, p) := -d_K(\mu) - \inf_{\substack{v(\cdot) \in L^2_\mu(R^d) \\ v(x) \in F(\mu, x)}} \int_{R^d} v(x).p(x)\, d\mu(x).$$

K is *invariant* iff the function $\mu \mapsto d_K(\mu)$ is a viscosity supersolution of

$$(L+2)u(\mu) + H_F^{inv}(\mu, D_\mu u(\mu)) = 0,$$

where, for all $\mu \in \mathcal{P}_2(R^d)$, $p \in L^2_\mu(R^d; R^d)$,

$$H_F^{inv}(\mu, p) := -d_K(\mu) - \sup_{\substack{v(\cdot) \in L^2_\mu(R^d) \\ v(x) \in F(\mu, x)}} \int_{R^d} v(x).p(x)\, d\mu(x).$$

where the relations $v(x) \in F(\mu, x)$ appaearring under the supremum and infimum means that for μ almost every $x \in R^d$ we have $v(x) \in F(\mu, x)$.

The notion of super solution has to be understood in viscosity sense namely : The functioon $u : \mathcal{P}_2(R^d) \mapsto \mathcal{P}_2(R^d)$ is a viscosity supersolution to

$$(L+2)u(\mu) + H(\mu, D_\mu u(\mu)) = 0,$$

if and only if for every $\bar{\mu} \in \mathcal{P}_2(R^d)$ and $\varepsilon > 0$ and any continuous differentiable [1] function $v : \mathcal{P}_2(R^d) \mapsto \mathcal{P}_2(R^d)$ such there exists $r > 0$ such that: $v(\bar{\mu}) = u(\bar{\mu})$ and

$$u(\nu) \le v(\nu) - \varepsilon W_2(\bar{\mu}, \nu) \quad \forall \nu \in \mathcal{P}_2(R^d) \text{ s.t. } W_2(\bar{\mu}, \nu) < r$$

we have

$$(L+2)v(\mu) + H(\mu, D_\mu v(\mu)) \le C\varepsilon,$$

for a constant which depends only on F.

The proof of the theorems is based on a comparison result for the Hamilton Jacobi equation that satisfy the following value functions defined on : $[0, T] \times \mathcal{P}_2(R^d)$

$$V^{viab}(t_0, \mu) := \inf_{(\mu_t)_{t \in [t_0, T]}} \int_{t_0}^{T} d_K(\mu_t)\, dt, \tag{4}$$

[1] For the definition of continouly differentiable function with Lions derivatives we refer te reader to Car (2018, 2013)

$$V^{inv}(t_0, \mu) := \sup_{(\mu_t)_{t \in [t_0, T]}} \int_{t_0}^{T} d_K(\mu_t)\, dt, \tag{5}$$

where the infimum and supremum are taken on the set of solutions of (2)-(3) with $\mu_{t_0} = \mu$.

REFERENCES

Ambrosio, L., Gigli N., Savaré G. (2008). *Gradient flows in metric spaces and in the space of probability measures.* Springer

Averboukh, Y. (2018). Viability theorem for deterministic mean field type control systems. *Set-Valued Var. Anal.*, 26, 993–1008.

Averboukh, Y., Marigonda A., Quincampoix M. (2021). Extremal shift rule and viability property for mean field-type control systems. *J. Optim. Theory Appl.* , 189 pp.244–270.

Badreddine, Z. Frankowska H.. (2022). Solutions to Hamilton-Jacobi equation on a Wasserstein space. *Calc. Var. Partial Differential Equations* , 61, 1–9.

Bonnet, B. Frankowska H.. (2021). Differential inclusions in Wasserstein spaces: the Cauchy-Lipschitz framework. *J. Differential Equations* , 271, 594–637.

Buckdahn R., Peng S., Quincampoix M., Rainer C. (1998). Existence of stochastic control under state constraints *omptes Rendus de l'Académie des Sciences. Série I. Mathématique*, 327, 17–22.

Cardaliaguet P., Quincampoix M. (2008). Deterministic differential games under probability knowledge of initial condition *Int. Game Theory Rev.* , 10, 1–16.

Cardaliaguet P., (2013). Notes on Mean Field Games *Unpublished Notes* .

Carmona R., Delarue F. (2018). *Probabilistic theory of mean field games with applications. I. & II Probability Theory and Stochastic Modelling.* Springer

Cavagnari G., Marigonda A., Quincampoix M. (2021). Compatibility of State Constraints and Dynamics for Multiagent Control Systems *J. Evol. Equ.* , 4, 4491-4537.

Gangbo W., Tudorascu A., On differentiability in the Wasserstein space and well-posedness for Hamilton-Jacobi equations *Journal de Mathématiques Pures et Appliquées,*, 125, pp.119–174.

Jimenez, C. Marigonda A.,, Quincampoix M. (2020). Optimal Control of Multiagent Systems in the Wasserstein Space *Calc. Var. Partial Differential Equations*, 59.

Jimenez, C. Marigonda A.,, Quincampoix M. (2021). Dynamical systems and Hamilton-Jacobi-Bellman equations on the Wasserstein space and their L^2 representations *Submitted.*

Marigonda A., Quincampoix M. (2021). Mayer control problem with probabilistic uncertainty on initial positions *J. Differential Equations,*, 264, 3212–3252.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

A semi-Lagrangian scheme for Hamilton-Jacobi-Bellman equations with oblique derivatives boundary conditions

E. Calzola [*] **E. Carlini .** [**] **X. Dupuis** [***] **F. J. Silva** [****]

[*] *"Sapienza", Università di Roma, Dipartimento di Matematica Guido Castelnuovo, 00185 Rome, Italy. (e-mail: calzola@mat.uniroma1.it)*
[**] *"Sapienza", Università di Roma, Dipartimento di Matematica Guido Castelnuovo, 00185 Rome, Italy. (e-mail: carlini@mat.uniroma1.it)*
[***] *Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université Bourgogne Franche-Comté, 21000 Dijon, France. (e-mail:xavier.dupuis@u-bourgogne.fr)*
[****] *Electrical Engineering Department, Institut de recherche XLIM-DMI, UMR 7252 CNRS, Faculté des Sciences et Techniques, (e-mail: francisco.silva@unilim.fr)*

Abstract: We investigate in this work a fully-discrete semi-Lagrangian approximation of second order possibly degenerate Hamilton-Jacobi-Bellman (HJB) equations on a bounded domain $\mathcal{O} \subset \mathbb{R}^N$ $(N = 1, 2, 3)$ with oblique derivatives boundary conditions. These equations appear naturally in the study of optimal control of diffusion processes with oblique reflection at the boundary of the domain.
The proposed scheme is shown to satisfy a consistency type property, it is monotone and stable. Our main result is the convergence of the numerical solution towards the unique viscosity solution of the HJB equation. The convergence result holds under the same asymptotic relation between the time and space discretization steps as in the classical setting for semi-Lagrangian schemes on $\mathcal{O} = \mathbb{R}^N$. We present some numerical results, in dimensions $N = 1$, 2, on unstructured meshes, that confirm the numerical convergence of the scheme.

Keywords: HJB equations oblique boundary conditions numerical approximation convergence analysis

1. INTRODUCTION

In this work we deal with the numerical approximation of the following parabolic Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{aligned} \partial_t u + H\left(t, x, Du, D^2u\right) &= 0 \quad \text{in } (0,T] \times \mathcal{O}, \\ L(t, x, Du) &= 0 \quad \text{on } (0,T] \times \partial\mathcal{O}, \\ u(0, x) &= \Psi(x) \quad \text{in } \overline{\mathcal{O}}. \end{aligned} \tag{1}$$

In the system above, $T > 0$, $\mathcal{O} \subset \mathbb{R}^N$ is a nonempty smooth bounded open set and H and L are nonlinear functions having the form

$$\begin{aligned} H(t, x, p, M) = \sup_{a \in A} \Big\{ &-\frac{1}{2} \operatorname{trace}\left(\sigma(t, x, a)\sigma(t, x, a)^\top M\right) \\ &- \langle \mu(t, x, a), p\rangle - f(t, x, a)\Big\}, \\ L(t, x, p) = \sup_{b \in B} &\left\{\langle \gamma(x, b), p\rangle - g(t, x, b)\right\}, \end{aligned}$$

where $A \subset \mathbb{R}^{N_A}$ and $B \subset \mathbb{R}^{N_B}$ are nonempty compact sets, $\sigma : [0,T] \times \overline{\mathcal{O}} \times A \to \mathbb{R}^{N \times N_\sigma}$, with $1 \le N_\sigma \le N$, $\mu : [0,T] \times \overline{\mathcal{O}} \times A \to \mathbb{R}^N$, $f : [0,T] \times \overline{\mathcal{O}} \times A \to \mathbb{R}$, $\gamma : \partial\mathcal{O} \times \mathcal{V} \to \mathbb{R}^N$, with $\mathcal{V} \subseteq \mathbb{R}^{N_B}$ being an open set containing B, $g : [0,T] \times \partial\mathcal{O} \times B \to \mathbb{R}$, and $\Psi : \overline{\mathcal{O}} \to \mathbb{R}$.

If $A = \{a\}$ and $B = \{b\}$, for some $a \in \mathbb{R}^{N_A}$ and $b \in \mathbb{R}^{N_B}$, and $\gamma(x, b) = n(x)$, with $n(x)$ being the unit outward normal vector to $\overline{\mathcal{O}}$ at $x \in \partial\mathcal{O}$, then (1) reduces to a standard linear parabolic equation with Neumann boundary conditions. In the general case, and after a simple change of the time variable in order to write (1) in backward form, the HJB equation (1) appears in the study of optimal control of diffusion processes with controlled reflection on the boundary $\partial\mathcal{O}$ (see e.g. Lions (1985) for the first order case, i.e. $\sigma \equiv 0$, and Lions (1983); Bouchard (2008) for the general case). Since the HJB equation (1) is possibly degenerate parabolic, one cannot expect the existence of classical solutions and we have to rely on the notion of viscosity solution (see e.g. Crandall et al. (1992)). Moreover, as it has been noticed in Lions (1982, 1985), in general the boundary condition in (1) does not hold in the pointwise sense and we have to consider a suitable weak formulation of it. We refer the reader to Lions (1985); Barles and Lions (1991) and Crandall et al. (1992); Barles (1993, 1999); Ishii and Sato (2004); Bourgoing (2008), respectively, for well-posedness results for HJB equations with oblique boundary condition in the first and second order cases.
The main purpose of this work is to provide a consistent,

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

stable, monotone and convergent SL scheme to approximate the unique viscosity solution to (1). By the results in Barles (1993), the latter is well-posed in $C([0,T]\times\overline{\mathcal{O}})$. Semi-Lagrangian schemes to approximate the solution to (1) when $\mathcal{O}=\mathbb{R}^N$ (see e.g. Camilli and Falcone (1995); Debrabant and Jakobsen (2013)) can be derived from the optimal control interpretation of (1) and a suitable discretization of the underlying controlled trajectories. These schemes enjoy the feature that they are explicit and stable under an inverse Courant-Friedrichs-Lewy (CFL) condition and, consequentely, they allow large time steps. A second important feature is that they permit a simple treatement of the possibly degenerate second order term in H. The scheme that we propose for $\mathcal{O}\neq\mathbb{R}^N$ preserves these two properties and seems to be the first convergent scheme to approximate (1) with the rather general assumptions. In particular, our results cover the stochastic and degenerate case. Consequently, from the stochastic control point of view, our scheme allows to approximate the so-called value function of the optimal control of a controlled diffusion process with possibly oblique reflection on the boundary $\partial\mathcal{O}$ (see Bouchard (2008)). The main difficulty in devising such a scheme is to be able to obtain a consistency type property at points in the space grid which are near the boundary $\partial\mathcal{O}$ while maintaining the stability. This is achieved by considering a discretization of the underlying controlled diffusion which suitably emulates its reflection at the boundary in the continuous case. We refer the reader to Milstein (1996) for a related construction of a semi-discrete in time approximation of a second order non-degenerate linear parabolic equation.

2. A SEMI-LAGRANGIAN SCHEME

Let $\Delta t>0$, set $N_{\Delta t}:=\lfloor T/\Delta t\rfloor$, $\mathcal{I}_{\Delta t}:=\{0,\dots,N_{\Delta t}\}$ and $\mathcal{I}^*_{\Delta t}:=\mathcal{I}_{\Delta t}\setminus\{N_T\}$. We define the time grid $\mathcal{G}_{\Delta t}:=\{t_k\,|\,t_k=k\Delta t,\ k\in\mathcal{I}_{\Delta t}\}$. Given $(k,i)\in\mathcal{I}^*_{\Delta t}\times\mathcal{I}_{\Delta x}$, $a\in A$, and $\ell=1,\dots,N_\sigma$, we define the discrete characteristics

$$y^{\pm,\ell}_{k,i}(a)=x_i+\Delta t\mu\left(t_k,x_i,a\right)\pm\sqrt{N_\sigma\Delta t}\sigma^\ell(t_k,x_i,a).$$

Let $\mathcal{I}=\{+,-\}\times\{1,\dots,N_\sigma\}$ and let $\bar c>0$ be a fixed constant. For any $\delta>0$ we set

$$(\partial\mathcal{O})_\delta:=\{x\in\mathbb{R}^N\,|\,d(x,\partial\mathcal{O})<\delta\}.$$

Tthere exist $R>0$ and two C^1 functions $(\partial\mathcal{O})_R\times B\ni(x,b)\mapsto p^{\gamma_b}(x)\in\partial\mathcal{O}$ and $(\partial\mathcal{O})_R\times B\ni(x,b)\mapsto d^{\gamma_b}(x)\in\mathbb{R}$, uniquely determined, such that

$$x=p^{\gamma_b}(x)+d^{\gamma_b}(x)\gamma_b(p^{\gamma_b}(x)),\quad\text{for all }(x,b)\in(\partial\mathcal{O})_R\times B.$$

Therefore, there exists $\overline{\Delta t}>0$ such that for all $\Delta t\in[0,\overline{\Delta t}]$, $(k,i)\in\mathcal{I}^*_{\Delta t}\times\mathcal{I}_{\Delta x}$, $a\in A$, $b\in B$, and $s\in\mathcal{I}$, the *reflected characteristic*

$$\tilde y^s_{k,i}(a,b):=\begin{cases}y^s_{k,i}(a) & \text{if } y^s_{k,i}(a)\in\overline{\mathcal{O}},\\ p^{\gamma_b}(y^s_{k,i}(a))-\bar c\sqrt{\Delta t}\gamma_b(p^{\gamma_b}(y^s_{k,i}(a))) & \text{else}\end{cases}$$

is well-defined. Let us also set

$$\tilde d^s_{k,i}(a,b):=\begin{cases}0 & \text{if } y^s_{i,k}(a)\in\overline{\mathcal{O}},\\ d^{\gamma_b}(y^s_{k,i}(a))+\bar c\sqrt{\Delta t} & \text{otherwise,}\end{cases}$$

$$\tilde g^s_{k,i}(a,b):=\begin{cases}0 & \text{if } y^s_{k,i}(a)\in\overline{\mathcal{O}},\\ g\left(t_k,p^{\gamma_b}\left(y^s_{k,i}(a)\right),b\right) & \text{otherwise.}\end{cases}$$

Notice that if $y^s_{k,i}(a)\notin\overline{\mathcal{O}}$, then

$$\tilde y^s_{k,i}(a,b)=y^s_{k,i}(a)-\tilde d^s_{k,i}(a,b)\gamma_b\left(p^{\gamma_b}(y^s_{k,i}(a))\right).$$

For $(k,i)\in\mathcal{I}^*_{\Delta t}\times\mathcal{I}_{\Delta x}$ and $\Phi:\mathcal{G}_{\Delta x}\to\mathbb{R}$, let us define $\mathcal{S}_{k,i}[\Phi]:A\times B\to\mathbb{R}$ by

$$\begin{aligned}\mathcal{S}_{k,i}[\Phi](a,b)\quad&:=\frac{1}{2N_\sigma}\sum_{s\in\mathcal{I}}\big[I[\Phi](\tilde y^s_{k,i}(a,b))\\&\qquad+\tilde d^s_{k,i}(a,b)\tilde g^s_{k,i}(a,b)\big]+\Delta t f(t_k,x_i,a)\end{aligned}$$

and set

$$S_{k,i}[\Phi]:=\inf_{a\in A,\,b\in B}\mathcal{S}_{k,i}[\Phi](a,b).$$

We consider the following fully discrete SL scheme to approximate the solution to (HJB).

$$\begin{aligned}U_{k,i}&=S_{k,i}\big[U_{k+1,(\cdot)}\big],\quad\text{for }(k,i)\in\mathcal{I}^*_{\Delta t}\times\mathcal{I}_{\Delta x},\\U_{N_{\Delta t},i}&=\Psi(x_i),\qquad\text{for } i\in\mathcal{I}_{\Delta x}.\end{aligned}$$

REFERENCES

Barles, G. (1993). Fully nonlinear Neumann type boundary conditions for second-order elliptic and parabolic equations. *J. Differential Equations*, 106(1), 90–106.

Barles, G. (1999). Nonlinear Neumann boundary conditions for quasilinear degenerate elliptic equations and applications. *J. Differential Equations*, 154(1), 191–224.

Barles, G. and Lions, P.L. (1991). Fully nonlinear Neumann type boundary conditions for first-order Hamilton-Jacobi equations. *Nonlinear Anal.*, 16(2), 143–153.

Bouchard, B. (2008). Optimal reflection of diffusions and barrier options pricing under constraints. *SIAM J. Control Optim.*, 47(4), 1785–1813.

Bourgoing, M. (2008). Viscosity solutions of fully nonlinear second order parabolic equations with L^1 dependence in time and Neumann boundary conditions. *Discrete Contin. Dyn. Syst.*, 21(3), 763–800.

Camilli, F. and Falcone, M. (1995). An approximation scheme for the optimal control of diffusion processes. *RAIRO Modél. Math. Anal. Numér.*, 29(1), 97–122.

Crandall, M.G., Ishii, H., and Lions, P.L. (1992). User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc. (N.S.)*, 27(1), 1–67.

Debrabant, K. and Jakobsen, E.R. (2013). Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. *Math. Comp.*, 82(283), 1433–1462.

Ishii, H. and Sato, M.H. (2004). Nonlinear oblique derivative problems for singular degenerate parabolic equations on a general domain. *Nonlinear Anal.*, 57(7-8), 1077–1098.

Lions, P.L. (1982). *Generalized solutions of Hamilton-Jacobi equations*, volume 69 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, Mass.-London.

Lions, P.L. (1983). Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. I. The dynamic programming principle and applications. *Comm. Partial Differential Equations*, 8(10), 1101–1174.

Lions, P.L. (1985). Neumann type boundary conditions for Hamilton-Jacobi equations. *Duke Math. J.*, 52(4), 793–820.

Milstein, G.N. (1996). Application of the numerical integration of stochastic equations for the solution of boundary value problems with Neumann boundary conditions. *Teor. Veroyatnost. i Primenen.*, 41(1), 210–218.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Learning Optimal Feedback Laws for Nonlinear Control Systems

Daniel Walter * **Karl Kunisch** **

* *Radon Institute, Austrian Academy of Sciences, A-4040 Linz, Austria (e-mail:daniel.walter@oeaw.ac.at)*
** *Institute of Mathematics and Scientific Computing, University of Graz, A-8010 Graz, Austria (e-mail:karl.kunisch@uni-graz.at).*

Abstract: A learning approach for optimal feedback gains for nonlinear continuous time control systems is proposed and analysed. Numerical results demonstrate the feasibilty of the approach, which allows to obtain suboptimal feedback gains, without focusing on directly solving the underlying Hamilton Jacobi Belman equation.

Keywords: Asymptotic stabilization, tracking, feedback control, learning techniques, neural networks.

1. INTRODUCTION

In this talk we focus on minimization problems associated to controlled dynamical systems of the form

$$\dot{y} = f(y) + g(y)u, \quad y(0) = y_0,$$

on the state space $\mathbb{R}^n$ and a time interval $I = [0, T]$ for $T \leq +\infty$. Here, $f\colon I \times \mathbb{R}^n \to \mathbb{R}^n$ describes the nonlinear dynamics of the system and y_0 denotes its initial state which is contained in a given compact set $Y_0 \subset \mathbb{R}^n$. We assume that the associated state trajectory y can be influenced by a control input $u(t) \in \mathbb{R}^m$ which enters the system via the, possibly state-dependent, control operator g. Our interest lies in choosing a control u^* such that the associated control-state pair (u^*, y^*) minimizes an energy $\mathcal{J}$,

$$\begin{cases} \inf\limits_{y,u} \ \mathcal{J}(y,u) := \displaystyle\int_0^T J(t, y(t), u(t)) \ \mathrm{d}t + \Psi_T(y(T)) \\ \quad s.t. \quad \dot{y} = f(y) + Bu, \quad y(0) = y_0, \end{cases} \tag{$P(y_0)$}$$

which is given in terms of a running cost functional $J\colon \mathbb{R}^n \times \mathbb{R}^m$ and a final time penalty $\Psi_T(y(T))$, in case $T < \infty$. Note that this general setting allows to consider stabilization problems on infinite time horizons, of the form

$$J(t, y, u) = \frac{1}{2}|Q_1 y|^2 + \frac{\beta}{2}|u|^2, \ \Psi_T(y) = 0, \ \text{if } T = +\infty,$$

as well as tracking-type optimal control problems with

$$J(t, y, u) = \frac{1}{2}|Q_1(y - y_d(t))|^2 + \frac{\beta}{2}|u|^2,$$

and

$$\Psi_T(y) = \frac{1}{2}|Q_2(y - y_d^T)|^2, \ \text{if } T < +\infty,$$

where $Q_1 \geq 0, Q_2 \geq 0, \beta > 0$, in a unified way.

2. OPTIMAL FEEDBACK CONTROL

Computing optimal controls by solving the *open loop* minimization problem comes with several drawbacks. First, such controls are only constructed as a function of time and thus cannot take into account possible perturbations of the dynamical system. Second, changing the initial condition y_0 the control action for a new initial condition requires to solve $P(y_0)$ all over.

For these reasons, we turn to optimal controls in feedback form, i.e. control inputs that are constructed as a function of the state variable at every time point t. In more detail we aim for a *feedback law* $F^*\colon \mathbb{R}^n \to \mathbb{R}^m$ such that:

- For every $y_0 \in Y_0$ there is a solution y^* to
$$\dot{y} = f(y) + g(y)F^*(y), \ y(0) = y_0.$$
- For every $y_0 \in Y_0$, the pair $(y^*, F^*(y^*))$ is a minimizer to $(P(y_0))$.

Constructing an optimal feedback F^* is closely related to the computation of the optimal value function

$$V^*(t, y_0) = \inf_{u,y} \int_t^T J(t, y(t), u(t)) \ \mathrm{d}t + \Psi_T(y(T))$$

which satisfies a *Hamilton-Jacobi-Bellman equation* (HJB), a partial differential equation on the state space. Once available, the optimal control to $(P(y_0))$ can be expressed in feedback form as $u^*(t) = -\frac{1}{\beta} g(t, y^*(t))^\top \nabla V^*(t, y^*(t))$. Following the HJB approach one is inevitably faced with the *curse of dimensionality*: If M degrees of freedom are used to discretize the HJB equation in each of the spatial directions, then this results in a discrete system with M^n degrees of freedom. Except for small dimensions n of the state equation this is unfeasible and alternatives must be sought.

2.1 A learning approach to feedback control

To circumvent the solution of the HJB-equation, in this talk we propose to replace the control u in $(P(y_0))$ by the closed loop expression $F_\theta^\varepsilon(y)$ where $\{F_\cdot^\varepsilon\}_{\varepsilon>0}$ denotes a family of parametrized models, each described by a parameter $\theta \in \mathcal{N}_\varepsilon \simeq \mathbb{R}^{N_\varepsilon}$, $N_\varepsilon \in \mathbb{N}$. We assume that $\{F_\cdot^\varepsilon\}_{\varepsilon>0}$ satisfies a $\mathcal{C}^1$-universal approximation property regarding the optimal feedback F^*:

Assumption 1. Suppose that there exists a compact set $N(Y_0)$ such that $y^*(t) \in N(Y_0)$ for every $t \in I$ and $y_0 \in Y_0$. The family $\{F^\varepsilon_\cdot\}_{\varepsilon>0}$ is said to be a universal approximator of F^* if there are parameters

$$\sup_{t\in I, y\in N(Y_0)} [|F^*(t,y) - F^\varepsilon_{\theta_\varepsilon}| + \|D_y F^*(t,y) - D_y F^\varepsilon_{\theta_\varepsilon}\|] \leq C\epsilon$$

for all $\varepsilon > 0$.

Under suitable assumptions on F^*, this requirement holds, for example for certain deep neural networks of varying width/depth or for piecewise polynomials of increasing total degree. Subsequently, an approximate optimal feedback law is determined from solving

$$\begin{cases} \inf\limits_{\theta} \int_0^T J(t, y(t), F^\varepsilon_\theta(t, y(t)))\ \mathrm{d}t + \Psi_T(y(T)) \\ \quad s.t. \quad \dot{y} = f(y) + g(y)F^\varepsilon_\theta(y), \quad y(0) = y_0, \end{cases} \tag{P_{Y_0}}$$

While it may be appealing to use a blackbox model for the approximation of the feedback law, we will show the merit of incorporating a priori knowledge on the structure of V^* (and its gradient) in the construction of the approximate feedback. For example, in the case of stabilizing a non-autonomous system with $f(0) = 0$ and $g(y) = B$, it is well known that the associated optimal feedback law is independent of time and satisfies $F^*(0) = 0$. Similarly, in the context of tracking-type optimal control problems, we can directly imprint the transversality condition

$$F^*(T, y) = -\frac{1}{\beta} g(y)^\top Q_2^\top Q_2 (y - y_d^T),$$

which follows from the necessary optimality conditions, onto the model. Finally, we could also exploit the connection between F^* and the gradient of the optimal value function V^*. This leads to an ansatz of the form

$$F^\varepsilon_\theta = -\frac{1}{\beta} g(\cdot)^\top \nabla V^\varepsilon_\theta(\cdot)$$

where V^ε_θ denotes a parametrized model for the scalar value function.

Of course, it can be expected that the effectiveness of this procedure depends on the location of the orbit $\mathcal{O} = \{y(t; y_0) : t \in (0, \infty)\}$ within the state space $\mathbb{R}^n$. To accommodate the case that $\mathcal{O}$ does not 'cover' the state-space sufficiently well, we propose to look at the ensemble of orbits departing from the compact set Y_0 of initial conditions and reformulate the problem accordingly. For this purpose we introduce a probability measure μ on Y_0 describing a "training set" of initial conditions and replace (P_{Y_0}) by

$$\left\{ \inf_{\theta} \int_{Y_0} \int_0^T J(t, \mathbf{y}(t), F_\theta(t, \mathbf{y}(t)))\ \mathrm{d}t + \Psi_T(\mathbf{y}(T)) d\mu(y_0) \right. \tag{$\mathcal{P}_\varepsilon$}$$

Here $\mathbf{y}$ is to be understood as an *ensemble* of state variables which assigns to every y_0 in the support of μ the solution of the closed loop state equation

$$\dot{y} = f(y) + g(y)F_\theta(y), \quad y(0) = y_0.$$

Our work gives mathematical rigor to this formulation including e.g. a discussion of its well-posedness.

Theorem 2. Under suitable conditions on f, g, Ψ, and appropriate constraints on the ensemble state $\mathbf{y}$ and the parameters θ, Problem $(\mathcal{P})$ admits at least one minimizing pair $(\mathbf{y}^*_\varepsilon, \theta^*_\varepsilon)$.

This requires a careful perturbation analysis of the underlying equation which is e.g. aggravated by the, potential, instability of the uncontrolled system as well as infinite time-horizons. For the practical realization of the approach we rely on first-order type methods. In this context, we provide a characterization of the gradient of the objective functional in the learning problem by means of adjoint calculus.

We also address the convergence of feedback laws obtained as the parametrized model gets more complex i.e. once $\varepsilon \to 0$. This leads to a variety of different convergence results depending on the particular type of problem under consideration (e.g. tracking-type vs. stabilization problems) as well as the training measure μ. In the aforementioned stabilization example and for an arbitrary training measure μ, we can show that the optimal objective functional values satisfy

$$\min(\mathcal{P}_\varepsilon) \to \int_{Y_0} V^*(y_0)\ \mathrm{d}\mu(y_0)$$

and thus approximate V^* in a suitable sense. Additionally, the optimal approximate ensemble states $\mathbf{y}^*_\varepsilon$ and the associated *approximate control actions* $BF^\varepsilon_{\theta^*_\varepsilon}(\mathbf{y}^*_\varepsilon)$ converge to the respective optimal quantities as $\varepsilon \to 0$. This is expressed in the following two theorems, for which the complete set of conditions on the problem data can be found in Walter (2021-1), Walter (2021-2), and Walter (2021-3).

Theorem 3. Assume that $\mathcal{P}(y_0)$ admits a unique optimal solution for every y_0 in the support of μ. Then there holds

$$(\mathbf{y}^*_\varepsilon, BF^\varepsilon_{\theta^*_\varepsilon}(\mathbf{y}^*_\varepsilon)) \to (\mathbf{y}^*, BF^*(\mathbf{y}^*))$$

where $\mathbf{y}^*$ is the optimal ensemble i.e. $\mathbf{y}^*(y_0) = y^*$ and convergence is obtained w.r.t a suitable topology on the space of ensemble functions.

In the important case of finite training data, the uniqueness assumption can be dropped and still convergence of the feedback controls can be obtained:

Theorem 4. Assume that $\mu = \sum_{j=1}^N \lambda_j \delta_{y_0^j}$, $y_0^j \in Y_0$. Then $\mathbf{y}^*_\varepsilon$ is of the form

$$\mathbf{y}^*_\varepsilon = (y^*_{\varepsilon,1}, y^*_{\varepsilon,2}, \ldots, y^*_{\varepsilon,N})$$

and every accumulation point of $\{(y^*_{\varepsilon,j}, F^\varepsilon_{\theta^*_\varepsilon}(y^*_{\varepsilon,j}))\}_{\varepsilon>0}$ is a minimizing pair of $(P(y_0^j))$.

Similar results can also be derived in the context of tracking-type problems with more general control operators. The cited papers Walter (2021-1), Walter (2021-2), and Walter (2021-3) contain numerical examples which illustrate the practical relevance of our learning approach. They range from highly unstable low dimensional systems to high dimensional examples stemming from the discretization of PDE systems. For related work with detailed numerical investigations we refer to Onken et al. (2021).

3. CONCLUSION

In summary, on the one hand, the results presented in this talk will show the great potential and success of learning feedback laws for optimal control problems. On the other hand, they also reveal open questions which stimulates further research. These include the development of fast and

reliable solution methods as well as the extension of our approach to PDE systems. Moreover, the approach itself is highly flexible in the sense that it directly allows to include control and/or state constraints into the problem as well as constraints on the feedback function itself.

REFERENCES

K. Kunisch, D. Walter. Semiglobal optimal feedback stabilization of autonomous systems via deep neural network approximation. ESAIM, Control Optim. Calc. Var., 27, 2021.

K. Kunisch, S. Rodrigues, D. Walter. Learning an optimal feedback operator semiglobally stabilizing semilinear parabolic equations. Appl. Math. Optim. 84, 2021.

K. Kunisch, D. Walter. Optimal feedback control of dynamical systems via function approximation. in preparation, 2022.

D. Onken, L. Nurbekyan, X. Li, S. W. Fung, S. Osher and L. Ruthotto. A Neural Network Approach for Real-Time High-Dimensional Optimal Control. arxive 2104.03270 (2021).

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Microscopic derivation of traffic flow models

P. Cardaliaguet * **N. Forcadel** **

* *CEREMADE, UMR CNRS 7534, Université Paris Dauphine-PSL, Place de Lattre de Tassigny, 75775 Paris Cedex 16, France. cardaliaguet@ceremade.dauphine.fr*
** *Normandie Univ, INSA de Rouen Normandie, LMI (EA 3226 - FR CNRS 3335), 76000 Rouen, France, 685 Avenue de l'Université, 76801 St Etienne du Rouvray cedex.*

Abstract: The goal of this note is to present a rigorous derivation of a macroscopic traffic flow model with a bifurcation or a local perturbation from a microscopic one. The microscopic model is a simple follow-the-leader with random parameters. The random parameters are used as a statistical description of the road taken by a vehicle and its law of motion. The limit model is a deterministic and scalar Hamilton-Jacobi equation on a network with a flux limiter, the flux-limiter describing how much the bifurcation or the local perturbation slows down the vehicles. The proof of the existence of this flux limiter relies on a concentration inequality and on a delicate derivation of a super-additive inequality.

Keywords: traffic flow, stochastic homogenization, Hamilton-Jacobi equations on networks.

1. INTRODUCTION

In this note, we present the results of Cardaliaguet et al. (2022) concerning the study of traffic flows models with a bifurcation consisting in a single incoming road which is divided after a junction into several outgoing ones. There are two main classes of models to describe these situations: microscopic models, which explain how each vehicle behaves in function of the vehicles in front; and macroscopic ones, taking the form of a conservation law in which the main unknown is the density of vehicles on the roads. Our aim is to start from simple microscopic models on a bifurcation and derive from these models continuous ones after scaling. The point is to get a better understanding of the continuous traffic flow models arising as the limit of discrete ones. Indeed there exists many different continuous models of traffic flow on a junction or with a local perturbation in the literature (see Garavello et al. (2006) and references therein) and the relation between these models is not completely clear. If the basic continuous model on a single straight road (the so-called LWR model, from Lighthill et al. (1955) and Richards (1956)) is well understood and justified by micro-macro limits in several contexts (see for example Di Francesco et al. (2015)), there is no consensus for problems with a junction or a bifurcation: the models are only obtained so far by heuristic arguments, with the exception of Forcadel et al. (2020) discussed below. Our goal is to show that the continuous model suggested in Imbert et al. (2017) pops up as the natural limit of follow-the-leader models. The continuous model in Imbert et al. (2017) takes the form of a flux limited Hamilton-Jacobi equation: it is a kind of integrated form of the basic LWR model outside the junction combined with a "flux limiting condition" on the junction. Our micro-macro derivation holds for a large class of follow-the-leader models, allowing for a possible heterogeneous behavior of the vehicles.

Our starting point is a microscopic model. Before describing it, let us recall that few discrete traffic flow models with a junction or a local perturbation exist in the literature (see for example Colombo et al. (2020); Andreianov et al (2018)). The only model proving micro-macro derivation in the case of a bifurcation is Forcadel et al. (2020) in which there are two outgoing roads and it is assumed (no too realistically) that every second vehicle takes a given road. In this setting the authors show that the convergence of the discrete problem to a flux limited solution of a Hamilton-Jacobi equation on a junction. One of the goals of the paper is to introduce a more realistic model in which one replaces the deterministic rule of Forcadel et al. (2020) by a random one (e.g., every second vehicle *in average* takes a given outgoing road). The introduction of randomness in traffic flow problems is natural and can be traced back to Chiabaut et al. (2010). The micro-macro derivation of the LWR model from a random one on a single road was established in Cardaliaguet et al. (2021). Here we present the corresponding result for a bifurcation.

2. MAIN RESULT

2.1 Short description of the microscopic model.

In our discrete model there is one incoming road and K outgoing ones, where $K \in \mathbb{N}$, $K \geq 1$. A position on the road is given by a pair (x, k) where x is a real number and k is a label in $\{0, \dots, K\}$. If x is nonpositive, then by convention $k = 0$ and the vehicle is on the incoming road. If x is positive then $k \in \{1, \dots, K\}$ and the vehicle is on the outgoing road k. The junction is an interval around $x = 0$, say, to fix the ideas, $[-R_0, 0]$. The vehicles are

labelled by $i \in \mathbb{Z}$. The position of the vehicle labelled i at time t is denoted by $U_i(t)$. The outgoing road the vehicle chooses is fixed from the beginning (independent of time) and denoted by $T_i \in \{1, \dots, K\}$. The motion of the vehicles is given by a leader-follower model: it satisfies the system of ordinary differential equations

$$\frac{d}{dt}U_i(t) = V_{Z_i}(U_{i+1}(t) - U_i(t), U_{\ell_i}(t) - U_i(t), U_i(t)), \qquad t \geq 0,\ i \in \mathbb{Z}, \quad (1)$$

where $V : \mathcal{Z} \times \mathbb{R}^3 \to \mathbb{R}_+$ is Lipschitz continuous in the three last variables (uniformly in the $z-$variable), nondecreasing with respect to the two middle ones and bounded by $\|V\|_\infty$. The type of the vehicle $i \in \mathbb{Z}$ is the random variable Z_i in $\mathcal{Z}$. We assume that $\mathcal{Z}$ is a finite set. Throughout the paper, $\Omega := \mathcal{Z}^{\mathbb{Z}}$ is endowed with the product $\sigma-$field $\mathcal{F}$ and with the product probability measure $\mathbb{P}$. We denote by $\tau : \mathbb{Z} \times \Omega \to \Omega$ the shift map defined by

$$(\tau_n \omega)_i = \omega_{i+n}, \qquad \forall \omega = (\omega_i)_{i \in \mathbb{Z}} \in \Omega,\ \forall n \in \mathbb{Z}.$$

We set $Z_i^\omega = \omega_i$ for $\omega = (\omega_i) \in \Omega$ and $i \in \mathbb{Z}$. As $\mathbb{P}$ is the product measure on Ω, this means that the $(Z_i)_i \in \mathbb{Z}$ are i.i.d.

The junction $\mathcal{R}$ is given by

$$\mathcal{R} = \bigcup_{k=0}^{K} \mathcal{R}^k$$

where

$\mathcal{R}^0 = (-\infty, 0] \times \{0\},\ \mathcal{R}^k = [0, +\infty) \times \{k\}$ for $k \in \{1, \dots, K\}$.

We also denote by $\overset{o}{\mathcal{R}}$ the interior of the roads:

$$\overset{o}{\mathcal{R}} = \bigcup_{k=0}^{K} \overset{o}{\mathcal{R}}{}^k,$$

where

$\overset{o}{\mathcal{R}}{}^0 = (-\infty, 0) \times \{0\},\ \overset{o}{\mathcal{R}}{}^k = (0, +\infty) \times \{k\}$ for $k \in \{1, \dots, K\}$.

We assume that all the vehicle are going or have gone through the junction and were ordered before going through the junction: $i+1$ is the label of the vehicle right in front of the vehicle i before this vehicle has gone through the junction. We denote by ℓ_i the label of the first vehicle in front of vehicle i taking the same outgoing road as i (in other words, $\ell_i = \inf\{j > i, T_i = T_j\}$). Each vehicle has a type Z_i encoding, on the one hand, the outgoing road the vehicle is taking or is going to take (namely, $T_i = T(Z_i)$ for a deterministic map $T : \mathcal{Z} \to \{1, \dots, K\}$) and, on the other hand, the "behavior" of the vehicle (for instance, if it is a truck or a race car). The velocity law $V = V_z(e_1, e_2, x)$ depends on the type $z \in \mathcal{Z}$ of the vehicle, the distances e_1 or e_2 to the next vehicle and the position x of the vehicle.

In order to obtain a limit model with a few unknowns and as simple as possible, we do not keep track of all the vehicles of a given type (in contrast with Colombo et al. (2018)). Instead we prefer a statistical description and assume that the types (Z_i) of the vehicles are random, independent and with the same law (i.i.d.); as a consequence the (T_i) are also i.i.d. In addition, we also suppose that the traffic is homogeneous outside the junction (see Assumption (H_3) below).

For later use we denote by $\pi^k := \mathbb{P}[T_i = k]$ the proportion of vehicle taking (or planning to take) road k.

2.2 The continuous macroscopic model.

The goal of the paper is to understand the behavior of the solution on large scale of time and space: namely, the behavior of $(x,t) \to \epsilon U_{[x/\epsilon]}(t/\epsilon)$, (where $[y]$ is the integer part of the real number y).

For $\epsilon > 0$, we look at the (scaled) traffic density of vehicles on each road:

$$m^\epsilon(dx, k, t) = \begin{cases} \epsilon \displaystyle\sum_{i \in \mathbb{Z},\ T_i = k} \delta_{\epsilon U_i(t/\epsilon)}(dx) \\ \qquad\qquad \text{if } x > 0,\ k \in \{1, \dots, K\} \\ \epsilon \displaystyle\sum_{i \in \mathbb{Z}} \delta_{\epsilon U_i(t/\epsilon)}(dx) \\ \qquad\qquad \text{if } x \leq 0,\ k = 0 \end{cases}$$

and want to understand the limit, as $\epsilon \to 0$, of m^ϵ. For this it is convenient to integrate in space m^ϵ and look instead at:

$$\nu^\epsilon(x, k, t) = \begin{cases} \epsilon(\pi^k)^{-1} \Big(\displaystyle\sum_{i \in \mathbb{Z},\ i \leq 0,\ T_i = k} \delta_{\epsilon U_i(t/\epsilon)}((x, +\infty)) \\ \qquad - \displaystyle\sum_{i \in \mathbb{Z},\ i > 0,\ T_i = k} \delta_{\epsilon U_i(t/\epsilon)}((-\infty, x]) \Big) \\ \qquad\qquad \text{if } x > 0,\ k \in \{1, \dots, K\} \\ \epsilon \Big(\displaystyle\sum_{i \in \mathbb{Z},\ i \leq 0} \delta_{U_i(t)}((x, +\infty)) \\ \qquad - \displaystyle\sum_{i \in \mathbb{Z},\ i > 0} \delta_{U_i(t)}((-\infty, x]) \Big) \\ \qquad\qquad \text{if } x \leq 0,\ k = 0. \end{cases} \quad (2)$$

Note that $\partial_x \nu^\epsilon = -m^\epsilon$ if $x \leq 0$ while $\partial_x \nu^\epsilon = -(\pi^k)^{-1} m^\epsilon$ if $x \geq 0$ and $k \in \{1, \dots, K\}$. This choice ensures the map ν^ϵ to be "almost continuous" at 0 since the vehicles are split between the K roads after the junction in proportion π^k for the road k. Our main result (Theorem 1) roughly states that, under suitable assumptions on V and if $\nu^\epsilon(\cdot, \cdot, 0)$ has a locally uniform (deterministic) limit $\nu_0(\cdot, \cdot)$ at time $t = 0$, then ν^ϵ has a.s. a locally uniform (deterministic) limit ν which is the unique viscosity solution to

$$\begin{cases} \partial_t \nu(x, k, t) + H^k(\partial_x \nu(x, k, t)) = 0 \quad \text{if } x \neq 0,\ t > 0 \\ \partial_t \nu + \\ \max\{\bar{A}, H^{0,+}(\partial_0 \nu), H^{1,-}(\partial_1 \nu), \dots, H^{K,-}(\partial_K \nu))\} = 0 \\ \qquad\qquad \text{at } x = 0 \\ \nu(x, k, 0) = \nu_0(x, k) \quad \text{for any } x, k, \end{cases} \quad (3)$$

where for $k \in \{0, \dots, K\}$, we denote by $H^{k,+}$ (resp. $H^{k,-}$) the largest nondecreasing (resp. nonincreasing) map below H^k.

The first equation is a Hamilton-Jacobi (HJ) equation in which the homogenized Hamiltonians $H^k(p)$ can be explicitly computed from the $\tilde{V}^k$. It corresponds to an integrated form of the LWR equation. The second equation describes the behavior of the vehicles at the junction (reduced after scaling to $x = 0$): we explain below the different terms. It roughly says that $\partial_t \nu + \bar{A} = 0$ at $x = 0$ (unless the HJ equation is satisfied at $x = 0$). The real number $\bar{A}$ is the so-called flux limiter. This is the main unknown of the paper. It quantifies how the traffic is slowed down by the junction. We show that

$$A_0 \le \bar{A} \le 0, \text{ where } A_0 := \max_{k\in\{0,\dots,K\}} \min_{p\in\mathbb{R}} H^k(p).$$

When $\bar{A} = A_0$, the flux is not limited at all. If $\bar{A} = 0$, then the traffic is completely stopped by the junction (this does not happen under our assumptions). The existence of $\bar{A}$ is the main point of the paper, which presents the first existence result of a flux limiter in the context of a stochastic homogenization problem. We show that $\bar{A}$ can be computed as follows:

$$\bar{A} = -\lim_{t\to+\infty} \frac{1}{t} \sharp\big\{i \in \mathbb{Z},\ \exists s \in [0,t],\ U_{e,i}(s) = 0\big\},$$

where $\sharp E$ denotes the number of elements of a set E, $e = (e^k)_{k=0,\dots,K}$ is such that $H^k(-1/e^k) = \min_p H^k(p)$ for any $k \in \{0,\dots,K\}$ and $(U_{e,i})$ is the solution to (1) with the "flat" initial condition $U_{e,i}(0) = e^k i$ (where $k = 0$ if $i \le 0$ and $k = T_i$ if $i \ge 0$). The quantity $\bar{A}$ can be interpreted as the maximal fraction of vehicles the junction can let pass given an amount of time. The introduction of Hamilton-Jacobi equations on a junction or stratified domains can be traced back to Achdou et al. (2013, 2015); Barles et al. (2013); Bressan et al. (2007); Imbert et al. (2013); Schieborn (2013); a general theory of flux limited solutions was developed in Imbert et al. (2017) (see also Barles et al. (2018)) with, as fundamental result, a comparison theorem; Lions et al. (2016, 2017) present different arguments for the comparison while Barles et al. (2018) proposes a general survey on the topic.

2.3 Assumptions

Let us state our standing assumptions on V_z:

- (H_1) For any $z \in \mathcal{Z}$, the map $(e_1, e_2, x) \to V_z(e_1, e_2, x)$ is Lipschitz continuous from $\mathbb{R}^2_+ \times \mathbb{R}$ to $\mathbb{R}_+$ and nondecreasing with respect to the first two variables;
- (H_2) There exists $e_{\max} > \Delta_{\min} > 0$ and $0 < R_2 < R_1 < R_0$, with $R_0 > e_{\max}$, such that for any $z \in \mathcal{Z}$, for any $(e_1, e_2, x) \in \mathbb{R}^2_+ \times \mathbb{R}$,
  - (i) $V_z(e_1, e_2, x) = 0$ if ($e_1 \le \Delta_{\min}$ and $x \le -R_2$) or if ($e_2 \le \Delta_{\min}$ and $x \ge -R_1$),
  - (ii) $V_z(e, e_2, x) = V_z(e_{\max}, e_2, x)$ and $V_z(e_1, e, x) = V_z(e_1, e_{\max}, x)$ if $e \ge e_{\max}$;
- (H_3) There exists $\tilde{V}^0, \dots, \tilde{V}^K : [0, +\infty) \to [0, +\infty)$ such that
$$V_z(e_1, e_2, x) = \begin{cases} \tilde{V}^0_z(e_1) \text{ if } x \le -R_0 \\ \tilde{V}^k_z(e_2) \text{ if } x \ge 0 \text{ and } T(z) = k. \end{cases}$$
- (H_4) For any $z \in \mathcal{Z}$ and any $k \in \{0,\dots,K\}$, there exists $h^k_{\max,z} \in (\Delta_{\min}, e_{\max}]$ such that $p \to \tilde{V}^k_z(p)$ is increasing and concave in $[\Delta_{\min}, h^k_{\max,z}]$ and constant on $[h^k_{\max,z}, +\infty)$;
- (H_5) There exists $\kappa > 0$ such that, for any $z \in \mathcal{Z}$,
  - (i) $V_z(e_1, e_2, x) = \tilde{V}^0_z(e_1)$ if $e_1 \le e_2$, $x \le -R_2$ and $V_z(e_1, e_2, x) \le \kappa$,
  - (ii) $\partial_x V_z(e_1, e_2, x) \ge 0$ if $x \in [-R_1, 0]$ and $V_z(e_1, e_2, x) \le \kappa$,
  - (iii) $V_z(e_1, e_2, x) > 0$ if $e_1 \wedge e_2 > \Delta_{\min}$.

Note that, by assumption (H_2), we have $\tilde{V}^k_z(e) = 0$ if $e \le \Delta_{\min}$ and $\tilde{V}^k_z(e) = \tilde{V}^k_z(e_{\max})$ if $e \ge e_{\max}$.

Some comment on the assumption are now in order. Assumption (H_2) is standard in the analysis of leader-follower models. The existence of $\Delta_{\min}$ prevents vehicles to collide (and could correspond to the size of the smallest vehicle for instance). The existence of $e_{\max}$ just says that the vehicles do not take into account the vehicles too far ahead. Assumption $R_0 > e_{\max}$ can be made without loss of generality. Assumption (H_3) means that the roads are homogeneous outside the bifurcation. This formalizes the fact that we concentrate here on a single bifurcation. Assumption (H_4) is also standard in the analysis of leader-follower models. There is one restriction though: the minimal distance such that the velocity has to be positive (i.e., here $\Delta_{\min}$) has to be the same for all vehicle and is not allowed to depend on the type of the vehicle; this restriction is related to the last (and technical) assumption (H_5). Assumption (H_5) has to do with the behavior of vehicles with slow velocity on the junction and ensures that the vehicles starting with a flat initial condition ($U_i(0) := e^k i)_{i\in\mathbb{Z}}$ (where $k = 0$ if $i \le 0$ and $k = T_i$ if $i \ge 0$ and e^k is such that $H^k(-1/e^k) = \min_p H^k(p)$) have a velocity bounded below by a positive constant independent of time and position. This last property is instrumental throughout the proofs. Assumption (H_5), without being unrealistic, is a little restrictive, but we do not know if it is possible to relax it.

2.4 The homogenized velocities and Hamiltonians.

Let $V^k_{\max,z} := \tilde{V}^k_z(h^k_{\max,z})$. Under assumptions (H1)—(H4), the map $\tilde{V}^k_z : [\Delta_{\min}, h^k_{\max,z}] \to [0, V^k_{\max,z}]$ is increasing and continuous for any $z \in \mathcal{Z}$ and any $k \in \{0,\dots,K\}$. We denote by $(\tilde{V}^k_z)^{-1}$ its inverse.

Let

$$\bar{v}^0 := \inf_{z\in\mathcal{Z}} \tilde{V}^0_z(e_{\max}), \qquad \bar{v}^k := \inf_{z\in\mathcal{Z},\ T(z)=k} \tilde{V}^k_z(e_{\max}). \quad (4)$$

We recall from Cardaliaguet et al. (2021) the definition of the homogenized velocities $\bar{V}^k$ and homogenized Hamiltonians: $\bar{V}^0$ is the inverse of the continuous increasing map defined on $(0, \bar{v}^0)$ by $v \to \mathbb{E}\left[(\tilde{V}^0_{Z_0})^{-1}(v)\right]$. We note that $\bar{V}^0$ is defined on $(\Delta_{\min}, \mathbb{E}\left[(\tilde{V}^0_{Z_0})^{-1}(\bar{v}^0)\right])$. We extend it for any $e \in [0, \Delta_{\min}]$ by $\bar{V}^0(e) = 0$ and for $e \ge \mathbb{E}\left[(\tilde{V}^0_{Z_0})^{-1}(\bar{v}^0)\right]$ by $\bar{V}^0(e) = \bar{v}^0$. In the same way we define $\bar{V}^k$ as the inverse of the continuous increasing map defined on $(0, \bar{v}^k)$ by $v \to \mathbb{E}\left[(\tilde{V}^k_{Z_0})^{-1}(v) \mid T_0 = k\right]$. It defines $\bar{V}^k$ on $(\Delta_{\min}, \mathbb{E}\left[(\tilde{V}^k_{Z_0})^{-1}(\bar{v}^0) \mid T_0 = k\right])$. We extend it for any $e \in [0, \Delta_{\min}]$ by $\bar{V}^k(e) = 0$ and for any $e \ge \mathbb{E}\left[(\tilde{V}^k_{Z_0})^{-1}(\bar{v}^0) \mid T_0 = k\right]$ by $\bar{V}^k(e) = \bar{v}^k$. The maps $\bar{V}^k$ (for $k \in \{0,\dots,K\}$) are continuous and bounded on $[0, +\infty)$.

We set, for any $k \in \{1,\dots K\}$,

$$\begin{cases} H^0(p) = p\bar{V}^0(-1/p), \\ H^k(p) = p\bar{V}^k(-1/(\pi^k p)), \end{cases} \quad p \in (-\infty, 0),$$
$$H^0(p) = H^k(p) = 0,\ \forall p \ge 0$$

and

$$A_0 = \max_{k\in\{0,\dots,K\}} \min_{p\in\mathbb{R}} H^k(p). \quad (5)$$

By Assumption $(H4)$, for $i \in \{0, \dots, K\}$, H^k is convex in $(-1/(\pi^k \Delta_{\min}), 0)$.

2.5 The main result

The main result of this note states that the system homogenizes: let $(U_i^{0,\epsilon})_{i\in\mathbb{Z}}$ be a deterministic family of initial conditions satisfying the compatibility condition: for any $i \in \mathbb{Z}$,

$$U_{i+1}^{0,\epsilon} \geq U_i^{0,\epsilon} + \Delta_{\min} \text{ if } U_{i+1}^{0,\epsilon} \leq -R_2$$

and

$$U_{\ell_i}^{0,\epsilon} \geq U_i^{0,\epsilon} + \Delta_{\min} \text{ for any } i \in \mathbb{Z}. \tag{6}$$

Up to relabel the indices, we also assume that $U_{i,0}^{\epsilon} \leq 0$ iff $i \leq 0$. Let U^ϵ be the solution of (1) with initial condition $(U_i^{0,\epsilon})_{i\in\mathbb{Z}}$.

Theorem 1. There is a set Ω_0 of full probability and a constant $\bar{A} < 0$ (the flux limiter) such that, if $(U_i^{0,\epsilon})_{i\in\mathbb{Z}}$ is a family of initial conditions such that the associated scaled function $\nu^\epsilon(\cdot,\cdot,0)$ defined by (2) (with $t = 0$) converges locally uniformly in $\mathcal{R}$ to a Lipschitz continuous map $\nu_0 : \mathcal{R} \to \mathbb{R}$, then, for any $\omega \in \Omega_0$, ν^ϵ converges locally uniformly in $\mathcal{R} \times [0, +\infty)$ to the unique continuous viscosity solution of the Hamilton-Jacobi equation (3) with flux limiter $\bar{A}$.

REFERENCES

Achdou, Y., Camilli, F., Cutrì, A., and Tchou, N. (2013). Hamilton-Jacobi equations constrained on networks. Nonlinear Differential Equations and Applications NoDEA, 20(3), 413-445.

Achdou, Y., and Tchou, N. (2015). Hamilton-Jacobi equations on networks as limits of singularly perturbed problems in optimal control: dimension reduction. Communications in Partial Differential Equations, 40(4), 652-693.

Andreianov, B., and Rosini, M. D. (2018). Microscopic selection of solutions to scalar conservation laws with discontinuous flux in the context of vehicular traffic. In Conference on Semigroups of Operators: Theory and Applications (pp. 113-135). Springer, Cham.

Barles, G., Briani, A., and Chasseigne, E. (2013). A Bellman approach for two-domains optimal control problems in $\mathbb{R}^N$. ESAIM: Control, Optimisation and Calculus of Variations, 19(3), 710-739.

Barles, G., Briani, A., Chasseigne, E., and Imbert, C. (2018). Flux-limited and classical viscosity solutions for regional control problems. ESAIM: Control, Optimisation and Calculus of Variations, 24(4), 1881-1906.

Barles, G., and Chasseigne, E. (2018). An illustrated guide of the modern approches of Hamilton-Jacobi equations and control problems with discontinuities. arXiv preprint arXiv:1812.09197.

Bressan, A., and Hong, Y. (2007). Optimal control problems on stratified domains. Networks & Heterogeneous Media, 2(2), 313.

Cardaliaguet, P., and Forcadel, N. (2021). From heterogeneous microscopic traffic flow models to macroscopic models. SIAM Journal on Mathematical Analysis, 53(1), 309-322.

Cardaliaguet, P., and Forcadel, N. (2022). Microscopic derivation of a traffic flow model with a bifurcation. Preprint

N. Chiabaut, L. Leclercq, and C. Buisson, From heterogeneous drivers to macroscopic patterns in congestion, Transportation Research Part B: Methodological, 44 (2010), pp. 299 – 308.

R.M. Colombo, C.Klingenberg, and M.-C. Meltzer. A multispecies traffic model based on the Lighthill-Whitham and Richards model, in Theory, Numerics and Applications of Hyperbolic Problems I, C. Klingenberg and M. Westdickenberg, eds., Cham, 2018, Springer International Publishing, pp. 375–394.

R. M. Colombo, H. Holden, and F. Marcellini, On the microscopic modeling of vehicular traffic on general networks. arXiv preprint arXiv:2002.09512v1

M. Di Francesco and M. D. Rosini, Rigorous derivation of nonlinear scalar conservation laws from follow-the-leader type models via many particle limit, Arch. Ration. Mech. Anal., 217 (2015), pp. 831–871.

Forcadel, N., and Salazar, W. (2020). Homogenization of a discrete model for a bifurcation and application to traffic flow. Journal de Mathmatiques Pures et Appliques, 136, 356-414.

M. Garavello and B. Piccoli, Traffic flow on networks, American institute of mathematical sciences Springfield, MO, USA, 2006.

Imbert, C., and Monneau, R. (2017). Flux-limited solutions for quasi-convex Hamilton-Jacobi equations on networks. In Annales scientifiques de l'Ecole normale suprieure, Vol. 50, No. 2, pp. 357-448.

Imbert, C., Monneau, R., and Zidani, H. (2013). A Hamilton-Jacobi approach to junction problems and application to traffic flows. ESAIM: Control, Optimisation and Calculus of Variations, 19(1), 129-166.

M. J. Lighthill and G. B. Whitham, On kinematic waves. ii. a theory of traffic flow on long crowded roads, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 229 (1955), pp. 317–345.

Lions, P. L., and Souganidis, P. E. (2016). Viscosity solutions for junctions: well posedness and stability. Rendiconti Lincei-matematica e applicazioni, 27(4), 535-545.

Lions, P. L., and Souganidis, P. E. (2017). Well-posedness for multi-dimensional junction problems with Kirchoff-type conditions. Rendiconti Lincei-Matematica e Applicazioni, 28(4), 807-816.

P. I. Richards, Shock waves on the highway, Operations research, 4 (1956), pp. 42–51.

Schieborn, D. (2006) Viscosity solutions of HamiltonJacobi equations of Eikonal type on ramified spaces. PhD thesis, Tbingen.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Time–dependent Hamilton–Jacobi equations on networks

Antonio Siconolfi

Dipartimento di Matematica, Sapienza Università di Roma, Piazzale A. Moro 5 Italy (e-mail: siconolf@mat.uniroma1.it).

Keywords: time–dependent Hamilton-Jacobi equations, Embedded networks, Viscosity solutions, Comparison principle, semidiscrete equations on graphs.

The purpose of this presentation is to study the well posedness of a time–dependent Hamilton–Jacobi equation, coupled with suitable additional conditions, posed on a network.

We consider a connected network Γ embedded in $\mathbb{R}^N$ with a finite number of arcs γ, which are regular simple curves parametrized in $[0,1]$, linking points of $\mathbb{R}^N$ called vertices, which make up a set we denote by $\mathbf{V}$. We define a Hamiltonian on Γ as a collection of Hamiltonians $H_\gamma : [0,1] \times \mathbb{R} \to \mathbb{R}$, indexed by arcs, with the crucial feature that Hamiltonians associated to arcs possessing different support, are totally unrelated.

The equations we deal with are accordingly of the form

$$u_t + H_\gamma(s, u') = 0 \qquad \text{in } (0,1) \times (0, +\infty)$$

on each arc γ, the aim being to uniquely select distinguished viscosity type solutions of each equation which can be assembled together continuously, making up a continuous function $u : \Gamma \times (0, +\infty) \longrightarrow \mathbb{R}$ with $u(\gamma(s), t)$ solution of the above equation for each γ. To accomplish it, one has to appropriately exploit the network geometry, via the adjacency condition between arcs and vertices, and the decisive issue for that is the right definition of supersolution. The subtle point in fact is that the conditions for supersolutions are not the same at all vertices, but are given taking into account the network structure.

The problem becomes discontinuous across all the one–dimensional interfaces of the form

$$\{(x,t), t \in [0, +\infty)\} \qquad \text{with } x \in \mathbf{V},$$

in contradiction to what happens for the stationary version of this kind of equations, where the discontinuities are located at the vertices, that is to say: they are finite and of zero dimension. This dimensional change explains why the analysis of evolutive equations on networks is by far more challenging than the stationary ones.

There are consequently few results available in the literature. A basic reference is (3) by Imbert and Monneau, where the topic is treated through PDE techniques, adapting tools from viscosity solutions theory, under the assumptions that the Hamiltonians in play are continuous, semiconvex and coercive. In (1), (2) applications of this theory are given.

We prove existence, uniqueness and stability of solutions on the network assuming convexity of the Hamiltonians, but without the growth conditions which allow applying Fenchel transform, so that an action functional cannot be defined. We do not have consequently representation formulae for solutions at hand, and our techniques employ purely PDE methods.

One of the main discoveries in (3) is that to get well posedness of the evolutive problem, the assignment of an initial datum at $t = 0$ is not enough. It must actually be coupled with a condition regarding the time derivative of solutions on the discontinuity interfaces. They qualify as *flux–limited* the corresponding solutions. We adopt here the same point of view, and the terminology of *flux limiter* as well.

Our definition of solution and the one of (3) are clearly the same outside the discontinuity interfaces, namely classical viscosity solutions. On the interfaces, the definition of subsolution coincides as well, while regarding supersolution, which is the most delicate point, the formulation is different, and our definition is stronger. However, a full comparison between the two notions cannot be done at present since the junctions considered in (3) have unbounded arcs while the arcs of our networks are bounded with two vertices as endpoints.

We think that our pattern is more related to the geometrical sense of the definition, and is more simple to write down, in particular because we take into account, for any arc, also the arc with the opposite orientation. This in particular implies that we do not have boundary vertices since any vertex has a least two adjacent arcs with opposite orientation. We would finally like to point out the fact that testing separately the equations on any arc can be a considerable advantage for a numerical analysis of the topic.

In contrast with (3), we do not need constructing special test functions at the vertices, and we do not use Crandall–Lions doubling variable method to get the comparison result.

Our method is different. We prove a comparison principle by associating the Hamilton–Jacobi equation to a semidiscrete problem posed on the discontinuity interfaces. This is the same road walked in (4), (5) for the stationary case, even if the evolutive setting brings in some complications.

The proof of the comparison result for the semidiscrete problem turns out to be quite simple, and it is then transferred to the initial equation exploiting the fundamental property that a continuous function $u : \Gamma \times [0, +\infty) \to \mathbb{R}$ is solution of the main problem if and only $u(\gamma(s), t)$ solves the HJ equations in the viscosity sense for any γ, and its trace on the discontinuity interfaces is solution of the semidiscrete problem.

A further relevant peculiarity of our techniques with respect to those in (3), is that we do not use special test functions at the vertices, more generally, we do not need functions testing at the same time solutions of equations with different Hamiltonians. For our definition, it is enough to consider viscosity test functions for the equations on the arcs, separately considered, plus test functions on the discontinuity interfaces. Finally, we do not use Perron–Ishii method to prove existence of solutions, but rely on a more constructive technique, showing first existence for small time interval and then gluing together the local solutions to get a solution global in time.

REFERENCES

[1] Guy Barles, Ariela Briani, Emmanuel Chasseigne, and Cyril Imbert. Flux-limited and classical viscosity solutions for regional control problems. *ESAIM Control Optim. Calc. Var.* 24 : 1881–1906, 2018.

[2] Giulio Galise, Cyril Imbert, and Régis Monneau. A junction condition by specified homogenization and application to traffic lights. *Analysis and PDE* 8 : 1891–1929, 2015.

[3] Cyril Imbert and Régis Monneau. Flux-limited solutions for quasi-convex Hamilton–Jacobi equations on networks. *Ann. Sci. Ec. Norm. Supér.* 50: 357–448, 2017.

[4] Marco Pozza and Antonio Siconolfi. Discounted Hamilton-Jacobi Equations on Networks and Asymptotic Analysis. *Indiana Un. Math J.* 70: pp. 1103–1129, 2021

[5] Antonio Siconolfi and Alfonso Sorrentino. Global Results for Eikonal Hamilton-Jacobi Equations on Networks *Analysis and PDE (11)* 1: 171–211, 2018.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

First Order Mean Field Games on Networks

Yves Achdou [*] **Paola Mannucci** [**] **Claudio Marchi** [***]
Nicoletta Tchou [****]

[*] *Université de Paris and Sorbonne Université, CNRS, Laboratoire Jacques-Louis Lions, (LJLL), F-75006 Paris, France, (e-mail: achdou@ljll-univ-paris-diderot.fr).*
[**] *Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Padova, Italy (e-mail: mannucci@math.unipd.it)*
[***] *Dipartimento di Ingegneria dell'Informazione & Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Padova, Italy, (e-mail: claudio.marchi@unipd.it)*
[****] *Université de Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, (e-mail: nicoletta.tchou@univ-rennes1.fr)*

Abstract: We study deterministic Mean Field Games with finite horizon in which the state space of the players is a network. In these games, the generic agent can control its dynamics: inside each edge, it can choose its velocity (which coincides with its control) and it can also choose, when it arrives at a vertex, the edge in which it enters. It will pay a cost which is formed by a running cost and a terminal cost; both these costs depend on the trajectory that it has chosen and on the evolution of the distribution of all agents. On the other hand, its position cannot affect the distribution of the whole population. As in the Lagrangian approach, we introduce a relaxed notion of Mean Field Games equilibria which relies on probability measures on trajectories on the network instead of probability measures on the network. Our main result is to establish the existence of such Mean Field equilibria.
With such an equilibrium at hand, we can introduce the value function and we prove that this function is a generalized solution to the associated first order Hamilton-Jacobi problem on the network.

Keywords: Control of constrained systems, Game theories, Generalized solutions of Hamilton-Jacobi equations, Hamiltonian trajectories in optimal control, Multi-agent systems, Optimal control theory.

1. INTRODUCTION

The theory of Mean Field Games studies the asymptotic behaviour of differential games (mainly in terms of their Nash equilibria) as the number of players tends to infinity. This theory started with the seminal papers by Lasry and Lions (2006a), Lasry and Lions (2006b), Lasry and Lions (2007) and by Huang et al. (2006). Models of this type have been intensively studied in the last decade; a detailed description of the achievements obtained in these years goes beyond the scope of this presentation. For a general overview we refer to the monographs by Achdou and Capuzzo Dolcetta (2010), Bensoussan et al. (2013), Cardaliaguet (2012) and Gomes et al. (2016).

In these games, the players are rational and indistinguishable and each one of them aims at choosing its own trajectory so to minimize its own cost which depends on the trajectory itself but also on the distribution of the whole population of agents. The player are "microscopic" and "identical": the position of a single player cannot affect the distribution of the whole population and the cost is the same for each one of them. The dynamics of the agents can be either stochastic or deterministic. Consider the case where the dynamics of the players are deterministic and the interaction among them is given by a nonlocal regularizing operator acting on the distribution of states of the agents. In the Euclidean setting, these models are described by a system of first order partial differential equations; indeed, the systems are formed by a continuity equation for the density of the distribution of the whole population (forward in time) and a Hamilton-Jacobi equation for the optimal value of a representative agent (backward in time), coupled with initial/final data (the initial distribution of the population and the final cost); see the paper by Cardaliaguet (2012).

We focus our attention on deterministic Mean Field Games, with finite horizon and a nonlocal regularizing interaction among the agents, in which the states of the agents are constrained in a network (in our setting, a network is given by a finite collection of vertices connected by continuous edges which cannot self-intersect). In these games, the generic agent can control its dynamics: inside each edge, it can choose its velocity (which coincides with its control) and it can also choose, when it arrives at a vertex, the edge in which it enters; in particular, it can also stop on any point of the network, either a vertex or a

point inside any edge. It will pay a cost which is the same for every agent and it is formed by a running cost and a terminal cost. The running cost is formed by two parts: a kinetic one (penalizing its high velocities) and a part depending on the chosen trajectory and on the evolution of the distribution of all agents; also the terminal cost depends on the distribution of all agents at final time. Moreover the costs (running and terminal) can change from edge to edge and other costs can appear for the times when the trajectories stay in the vertices. All the costs depend on the distribution of agents in a nonlocal regularizing manner. The position of a single agent cannot affect the distribution of the whole population; in other words, if an agent knows the evolution of the distribution of the whole population, then it has only to choose its trajectory so to minimize its cost.

Clearly these problems are encompassed in the framework of deterministic state constrained Mean Field Games. On one hand, even the study of deterministic control problems on networks or other irregular sets is rather recent (see: Achdou et al. (2013), Barles et al. (2014), Imbert et al. (2013), Imbert and Monneau (2017), Lions and Souganidis (2016) and Morfe (2020)) and this topic still displays a lot of interesting open problems. On the other hand, it is worth to recall that an important issue for these Mean Field Games is that the agents could concentrate on the boundary of the constraint (namely, in the vertices of our network). Indeed we provide an example where the distribution of agents develops singular measures immediately after the initial time. This issue makes difficult to characterize the state distribution by means of a partial differential equation. In order to overcome this difficulty, we shall follow the approach introduced by Cannarsa and Capuani (2018) for deterministic Mean Field Games constrained in the closure of a regular bounded open set. This approach is based on the Lagrangian setting rather than in terms of a system of differential equations.

As in the Lagrangian approach (see also Benamou and Brenier (2000), Benamou and Carlier (2015) and Cardaliaguet et al. (2016)), we shall describe the evolution of the game in terms of a probability measures on the set of admissible trajectories on the network instead of a probability measure on the network. Roughly speaking, to each probability measure on the set of admissible trajectories, we associate a time-dependent family of probabilities on the network which, in turns, permits to define a running cost and a terminal cost. A *Mean Field equilibrium* is a probability on the set of admissible trajectories whose support is contained in the set of optimal trajectories for the cost associated to that probability.

Our main result is the existence of such a Mean Field equilibrium. The proof is based on the application of Kakutani fixed point theorem. To this end, we shall need some properties of optimal control problems on networks with finite horizon: existence of optimal trajectories and an approximation result for trajectories on the network.

With such an equilibrium at hand, the costs for the agents are well defined and, consequently a value function can be introduced. We also prove that this value function is a generalized solution to the associated first order Hamilton-Jacobi problem on the network.

REFERENCES

Achdou, Y., Camilli, F., Cutrì, A., and Tchou, N. (2013). Hamilton-Jacobi equations constrained on networks. *NoDEA Nonlinear Differential Equations Appl.*, 20, 413–445.

Achdou, Y. and Capuzzo Dolcetta, I. (2010). Mean field games: numerical methods. *SIAM J. Numer. Anal.*, 48, 1136–1162.

Barles, G., Briani, A., and Chasseigne, E. (2014). A Bellman approach for regional optimal control problems in r^n. *SIAM J. Control Optim.*, 52, 1712–1744.

Benamou, J.D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84, 375–393.

Benamou, J.D. and Carlier, G. (2015). Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167, 1–26.

Bensoussan, A., Frehse, J., and Yam, P. (2013). *Mean Field Games and Mean Field Type Control Theory*. Springer Briefs in Mathematics, Springer, New York.

Cannarsa, P. and Capuani, R. (2018). Existence and uniqueness for mean field games with state constraints. In P. Cardaliaguet, A. Porretta, and F. Salvarani (eds.), *PDE models for multi-agent phenomena*, volume 28, 49–71. Springer Indam Ser., Cham.

Cardaliaguet, P. (2012). Notes on mean field games. *available online*.

Cardaliaguet, P., Mészáros, A., and Santambrogio., F. (2016). First order mean field games with density constraints: pressure equals price. *SIAM J. Control Optim.*, 54, 2672–2709.

Gomes, D., Pimentel, E., and Voskanyan, V. (2016). *Regularity Theory for Mean-Field Game Systems*. Springer Briefs in Mathematics, Springer, Berlin.

Huang, M., Malhamé, R., and Caines, P. (2006). Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.*, 6, 221–251.

Imbert, C. and Monneau, R. (2017). Flux-limited solutions for quasi-convex Hamilton-Jacobi equations on networks. *Ann. Sci. Éc. Norm. Supér. (4)*, 50, 357–448.

Imbert, C., Monneau, R., and Zidani, H. (2013). A Hamilton-Jacobi approach to junction problems and application to traffic flows. *ESAIM Control Optim. Calc. Var.*, 19, 129–166.

Lasry, J. and Lions, P.L. (2006a). Jeux à champ moyen. I. le cas stationnaire. *C. R. Math. Acad. Sci. Paris*, 343, 619–625.

Lasry, J. and Lions, P.L. (2006b). Jeux à champ moyen. II. horizon fini et contrôle optimal. *C. R. Math. Acad. Sci. Paris*, 343, 679–684.

Lasry, J. and Lions, P.L. (2007). Mean field games. *Jpn. J. Math.*, 2, 229–260.

Lions, P.L. and Souganidis, P. (2016). Viscosity solutions for junctions: well posedness and stability. *Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl.*, 27, 535–545.

Morfe, P. (2020). Convergence and rates for Hamilton-Jacobi equations with Kirchoff junction conditions. *NoDEA Nonlinear Differential Equations Appl.*, 27, no. 10.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Deterministic mean field games with non coercive Hamiltonian

Paola Mannucci * **Claudio Marchi** ** **Nicoletta Tchou** ***

* *Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, mannucci@math.unipd.it*
** *Dipartimento di Ingegneria dell'Informazione & Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, claudio.marchi@unipd.it*
*** *Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, nicoletta.tchou@univ-rennes1.fr*

Abstract: We study some models of evolutive deterministic mean field games with finite time horizon where the Hamiltonian is not coercive in the gradient term because the dynamic of the generic player has some forbidden directions. We study the existence of weak solutions and their representation by means of relaxed equilibria in the Lagrangian setting which are described by a probability measure on optimal trajectories.

Keywords: Mean Field Games, first order Hamilton-Jacobi equations, continuity equation, Fokker-Planck equation, noncoercive Hamiltonian.

1. INTRODUCTION

We study evolutive first order mean field games (MFG) systems where the Hamiltonian is neither strictly convex nor coercive. They are described by a system of PDEs coupling a continuity equation for the density of the distribution of the states (forward in time) and a Hamilton-Jacobi equation for the optimal value of a representative agent (backward in time).
We consider systems in $R^n \times (0,T)$ of the form

$$\begin{cases} (i) \quad -\partial_t u + \dfrac{|D_{\mathcal{H}} u|^2}{2} = F[m(t)](x) \\ (ii) \quad \partial_t m - div_{\mathcal{H}}(m D_{\mathcal{H}} u) = 0 \\ (iii) \quad m(x,0) = m_0(x), u(x,T) = G[m(T)](x), \end{cases} \tag{1}$$

where $D_{\mathcal{H}}$ and $div_{\mathcal{H}}$ are respectively the so called *horizontal gradient* and the *horizontal divergence*, i.e. the gradient and the divergence along n_0 prescribed vector fields, with $n_0 \leq n$. More precisely, we consider n_0 vector fields $X_i(x)$, $i = 1, \cdots, n_0$ in R^n and we call $B(x)$ the $n \times n_0$ matrix whose columns are X_i. Using the matrix B we can write:

$$D_{\mathcal{H}} u = DuB, \; div_{\mathcal{H}}(m D_{\mathcal{H}} u) = div(mDuBB^T).$$

These MFG systems arise when the generic player can move in the whole space but it must follow *horizontal curves* with respect to the vector fields X_i :

$$x'(s) = B(x(s))\alpha(s), \quad x(t) = x \tag{2}$$

where $\alpha \in R^{n_0}$. Each agent wants to choose the control α in $L^2([t,T]; R^{n_0})$ in order to minimize the cost

$$J^m_{x,t}(\alpha) := \int_t^T \left[\frac{1}{2}|\alpha(\tau)|^2 + F[m(\tau)](x(\tau))\right] d\tau + G[m(T)](x(T)) \tag{3}$$

where $m(\cdot)$ is the evolution of the whole population's distribution while $(x(\cdot), \alpha(\cdot))$ is a trajectory obeying to (2).

We suppose that the coefficients of the matrix B have at most a linear growth with respect to the space variable x.

Let us observe three important issues of these MFG systems: (i) the Hamiltonian $H(x,p) = \frac{1}{2}|p\,B(x)|^2$ is not coercive in p, (ii) the system is in the whole space, (iii) in equation (1)-(ii) the coefficient of the first order term may have quadratic growth in x.
Point (i) prevents the application of standard approaches for first order MFG (for instance, see (BFY; C)) because they require uniform coercivity of the Hamiltonian.

On the other hand, points (ii) and (iii) give rise to some difficulties for applying the vanishing viscosity method, especially for the Cauchy problem for equation (1)-(ii) with the viscosity term. Actually in this problem the coefficients grow "too much at infinity" and one cannot invoke nor standard results for the well-posedness of the problem neither its interpretation in terms of a stochastic optimal control problem.

We get two results: the former one is to prove the existence of a weak solution to system (1) while the latter, and main, one is to prove that this weak solution is also a *mild* solution in the sense introduced by Cannarsa and Capuani (CC) for the case of state-constrained MFG where the agents control their velocity. Roughly speaking, as in

⋆ The first and the second authors were partially supported by GNAMPA-INdAM by the research project of the University of Padova "Mean-Field Games and Nonlinear PDEs", by the Fondazione CaRiPaRo Project "Nonlinear Partial Differential Equations: Asymptotic Problems and Mean-Field Games" and by KAUST project OSR-2017-CRG6-3452.01. The third author has been partially funded by the ANR project ANR-16-CE40-0015-01.

the Lagrangian approach for MFG, this property means that, for a.e. starting state, the agents follow optimal trajectories for the optimal control problem associated to the Hamilton-Jacobi equation.

In order to obtain the existence of a weak solution, we establish several properties of the solution to the Hamilton-Jacobi equation (1)-(i) (as semiconcavity, Lipschitz continuity, regularity of the optimal trajectories for the associated optimal control problem). Afterwards, we adapt the techniques introduced by PL. Lions in his lectures at Collège de France, see (C), and also (AMMT; AMMT2; MMMT) for similar approaches for some noncoercive Hamiltonians). To get the result we perform three approximations: a completion B^ε of B, a vanishing viscosity procedure with the *Euclidean* Laplacian and a truncation argument of the coefficients of matrix B. The completion B^ε fulfills $\det B^\varepsilon (B^\varepsilon)^T \neq 0$ for any $x \in R^n$ which is a crucial property for getting uniqueness of optimal trajectory for m_0-a.e. starting point. The vanishing viscosity procedure permits to exploit the regularity results of the Laplacian while the truncation argument permits to avoid parabolic Cauchy problems with coefficients growing "too much" at infinity.
Finally, we shall prove that this weak solution is also a *mild* solution in the sense introduced in (CC). In order to prove that our solution is in fact a mild solution, we shall use the superposition principle obtained in (AGS).

It is worth noting that our techniques relies on some compactness of initial distribution of players and on sublinear growth of the coefficients of B.

Uniqueness holds under classical hypothesis on the monotonicity of F and G as in (C).

Some examples Our result can be applied to the following examples:

– Completely degenerate case. In the state space R^n:

$$B(x) = \begin{pmatrix} I_{n_0} & 0_{n_0,(n-n_0)} \\ 0_{(n-n_0),n_0} & 0_{(n-n_0),(n-n_0)} \end{pmatrix}$$

where I_i is the identity matrix $i \times i$ while $0_{i,j}$ is the null matrix $i \times j$, $i, j = n_0, n-n_0$. With this matrix, the generic player in the MFG controls only its first n_0 coordinates.
For example for $n = 2$ and $n_0 = 1$ the dynamics and the Hamiltonian assume the following form

$$x_1' = \alpha_1,\ x_2' = 0.$$
$$H(x,p) = \frac{1}{2}p_1^2 = \frac{|pB(x)|^2}{2}.$$

– Grushin case. In the state space R^2:

$$B(x) = \begin{pmatrix} 1 & 0 \\ 0 & x_1 \end{pmatrix}.$$

In this case the dynamics and the Hamiltonian assume the following form

$$x_1' = \alpha_1,\ x_2' = x_1\alpha_2.$$
$$H(x,p) = \frac{1}{2}(p_1^2 + (x_1p_2)^2) = \frac{|pB(x)|^2}{2}.$$

– Heisenberg case. In the state space R^3:

$$B(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -x_2 & x_1 \end{pmatrix}.$$

In this case the dynamics and the Hamiltonian assume the following form

$$x_1' = \alpha_1,\ x_2' = \alpha_2,\ x_3' = -x_2\alpha_1 + x_1\alpha_2.$$
$$H(x,p) = \frac{1}{2}((p_1 - x_2p_3)^2 + (p_2 + x_1p_3)^2) = \frac{|pB(x)|^2}{2}.$$

REFERENCES

[AMMT] Y. Achdou, P. Mannucci, C. Marchi, N. Tchou. Deterministic mean field games with control on the acceleration. *NoDEA Nonlinear Differential Equations Appl.*, 27 no. 3, p. 33, (2020).

[AMMT2] Y. Achdou, P. Mannucci, C. Marchi, N. Tchou. Deterministic Mean Field Games with Control on the Acceleration and State Constraints *SIAM J. Math. Anal.*, 54, no. 3, 3757?3788, (2022).

[AGS] L. Ambrosio, N. Gigli and G. Savaré, Gradient flows in metric spaces and in the space of probability measures, *Lectures in Mathematics ETH Zürich.* Birkhaüser Verlag, Basel 2005.

[BFY] A. Bensoussan, J. Frehse, P. Yam, Mean field games and mean field type control theory, Springer Briefs in Mathematics. Springer, New York, 2013.

[CC] P. Cannarsa, R. Capuani, Existence and uniqueness for Mean Field Games with state constraints, *PDE models for multi-agent phenomena*, 49–71, Springer INdAM Ser., 28, Springer, Cham, 2018.

[C] P. Cardaliaguet, Notes on Mean Field Games, from P.L. Lions lectures at College de France (2012), available at https://www.ceremade.dauphine.fr/cardalia/MFG20130420.pdf.

[MMT] P. Mannucci, C. Marchi, N. Tchou, Non-coercive unbounded first order Mean Field Games: the Heisenberg example, *J. Differential Equations*, 309 , 809-840- 2022.

[MMMT] P. Mannucci, C. Mariconda, C. Marchi, N. Tchou, Non-coercive first order Mean Field Games, *J. Differential Equations*, 269 no. 5, 4503–4543, 2020.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Carlemen Linearization of Nonlinear Systems and Their Convergent Finite-Section Approximations ⋆

Arash Amini * **Cong Zheng** ** **Qiyu Sun** ** **Nader Motee** *

* *Department of Mechanical Engineering and Mechanics, Lehigh University, Bethlehem, PA 18015 (e-mail:ara416@lehigh.edu, motee@lehigh.edu).*
** *Department of Mathematics, University of Central Florida, Orlando, Florida 32816 (e-mail:acongz@Knights.ucf.edu, qiyu.sun@ucf.edu).*

Abstract: In his 1932 paper, Carleman proposes a linearization method to transform a given finite-dimensional nonlinear system defined by an analytic function into an equivalent infinite-dimensional linear system with (usually) unbounded operators. Finite truncation of the transformed system has been used to study dynamical properties, learning, and control of such nonlinear systems. One of the fundamental problems in this context is to quantify the effectiveness of such finitely truncated models. In this paper, we provide explicit error bounds and prove that the trajectory of the truncated system stays close to that of the original nonlinear system over a quantifiable time interval. This is particularly important in several applications, including Model Predictive Control, to choose proper truncation lengths for a given sampling period and employ the resulting truncated system as a good approximation of the nonlinear system.

Keywords: Application of nonlinear analysis and design, Model reduction, Lifting Operators, Carleman Linearization, Nonlinear System

1. INTRODUCTION

Almost all natural, physical, and engineered systems are time-varying and nonlinear, and often they need to be modeled using partial differential equations. The nonlinear nature of models has imposed fundamental challenges for the analysis and design of real-world systems. Some traditional design methods rely on a linear system obtained from the first-order approximation of the nonlinear system's right-hand side. To study the properties of a nonlinear system, researchers have developed several frameworks over the past century Koopman (1931); Carleman (1932); Isidori (2013); Arnold et al. (1988); Wiggins et al. (1990); Khalil (2002). One of the mainstream approaches, which has attracted researchers' attention for decades, represents a finite-dimensional nonlinear system as an infinite-dimensional linear system using lifting operators. Carleman linearization Carleman (1932) and Koopman operator Koopman (1931) are two of the most prominent examples closely connected in spirit. Carleman linearization is a procedure, also referred to as lifting, that transforms a finite-dimensional nonlinear system into an infinite-dimensional linear system Kowalski and Steeb (1991); Carleman (1932). Several follow-up works have tried to address various aspects of this method Bellman and Richardson (1963); Brockett (2014); Steeb and Wilhelm (1980); Bertsekas (2011). Albeit the lifting procedure results in a linear system, one usually has to deal with systems with unbounded operators and their associated convergence issues.

Although Carleman linearization appears to be very appealing for analyzing and controlling nonlinear systems, one should take extra care when working with unbounded infinite-dimensional matrices. Unless there are some useful structures, handling systems with unbounded infinite-dimensional matrices is exceptionally challenging. A common remedy is to truncate the infinite-dimensional system and then utilize the truncated system for analysis and control purposes. Several stories reported about the successful employment of Carleman linearization in the control systems community. The author of Banks (1992) identifies relationships between Carleman linearization and Lie series and then utilizes it to design optimal control laws for infinite-dimensional systems. In Svoronos et al. (1994), a straightforward method for discretizing nonlinear continuous-time systems via Carleman approximation is provided. By exploiting the lifted system's structure, the authors of Amini et al. (2020a,b) propose an efficient method to quadratize and solve the Hamilton-Jacobi-Bellman (HJB) through an exact iterative method.

Despite numerous Carleman linearization applications in nonlinear control systems, the convergence of finite truncation of the lifting system was not addressed for general nonlinear systems. Reference Forets and Pouly (2017) finds some approximation bounds for the class of polynomial systems, i.e. systems whose right-hand sides are finite order polynomials. To the best of our knowledge, our

⋆ This work was supported in parts by the AFOSR FA9550- 19-1-0004 and ONR N00014-19-1-2478, and NSF DMS-1816313.

work is the first to quantify error bounds for the general class of time-varying nonlinear systems whose right-hand sides are analytic. We prove that the first block of the truncated system (which provides an estimate of the actual state) converges exponentially fast to the original system's solution when the order of truncation increases. In this extended abstract we only included our main result and removed the proofs due to the page limitation.

2. BACKGROUND: CARLEMAN LINEARIZATION

We consider the class of nonlinear systems whose dynamics are governed by

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}) \tag{2.1}$$

for all $t \geq t_0$ with a nonzero initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$, at a neighborhood of its equilibrium at the origin, where the state of the system is denoted by $\mathbf{x} = [x_1, \ldots, x_d]^T \in \mathbb{R}^d$ and the components of function

$$\mathbf{f}(t, \mathbf{x}) = \big[f_1(t, \mathbf{x}), \ldots, f_d(t, \mathbf{x})\big]^T,\ t \geq t_0$$

are analytic on a neighborhood of the equilibrium. Without loss of generality, it is assumed that $\mathbf{f}(t, \mathbf{0}) = \mathbf{0}$. For a given multi-index vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^T \in \mathbb{Z}_+^d$, let us define $\mathbf{x}_{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ and express the Maclaurin series expansion of the vector-valued function $\mathbf{f}(t, \mathbf{x})$ by

$$\mathbf{f}(t, \mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{Z}_+^d \setminus \{\mathbf{0}\}} \mathbf{f}_{\boldsymbol{\alpha}}(t) \mathbf{x}_{\boldsymbol{\alpha}} \tag{2.2}$$

in which $\mathbf{f}_{\boldsymbol{\alpha}}(t) = [f_{1,\boldsymbol{\alpha}}(t), \ldots, f_{d,\boldsymbol{\alpha}}(t)]^T$. The conventional Carlemen linearization of the nonlinear dynamic system (2.1) starts from the reformulation

$$\dot{x}_j = \sum_{\alpha \in \mathbb{Z}_+^d \setminus \{\mathbf{0}\}} f_{j,\alpha}(t)\, \mathbf{x}_{\boldsymbol{\alpha}} \tag{2.3}$$

for every $j = 1, \ldots, d$. The standard Euclidean basis for $\mathbb{R}^d$ is denoted by $\mathbf{e}_j = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ for $j = 1, \ldots, d$ and it is assumed that

$$f_{j,\boldsymbol{\alpha}} = 0 \ \text{ if } \ \boldsymbol{\alpha} \notin \mathbb{Z}_+^d \setminus \{\mathbf{0}\}. \tag{2.4}$$

From (2.3), the derivative of monomial $\mathbf{x}_{\boldsymbol{\alpha}}$ for every $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d] \in \mathbb{Z}_+^d \setminus \{\mathbf{0}\}$, can be calculated as

$$\begin{aligned}\dot{\mathbf{x}}_\alpha &= \sum_{j=1}^{d} \alpha_j\, \mathbf{x}_{\boldsymbol{\alpha}-\mathbf{e}_j} \dot{x}_j = \sum_{j=1}^{d} \alpha_j \mathbf{x}_{\boldsymbol{\alpha}-\mathbf{e}_j} \sum_{\boldsymbol{\gamma} \in \mathbb{Z}_+^d \setminus \{\mathbf{0}\}} f_{j,\boldsymbol{\gamma}}(t) \mathbf{x}_{\boldsymbol{\gamma}} \\ &= \sum_{\boldsymbol{\beta} \in \mathbb{Z}_+^d \setminus \{\mathbf{0}\}} \left(\sum_{j=1}^{d} \alpha_j f_{j,\boldsymbol{\beta}-\boldsymbol{\alpha}+\mathbf{e}_j}(t) \right) \mathbf{x}_{\boldsymbol{\beta}}\end{aligned}$$

with initial condition $\mathbf{x}_{\boldsymbol{\alpha}}(t_0) = \mathbf{x}_{\boldsymbol{\alpha}}^0 = x_1^{\alpha_1}(t_0) \cdots x_d^{\alpha_d}(t_0)$. Let us define $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$ for $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^T \in \mathbb{Z}_+^d$ and

$$\mathbb{Z}_k^d = \Big\{ \boldsymbol{\alpha} \in \mathbb{Z}_+^d \ \Big|\ |\boldsymbol{\alpha}| = k \Big\}$$

for $k \geq 0$. Regrouping all $\boldsymbol{\alpha} \in \mathbb{Z}_k^d$ for every $k \geq 1$ and defining $\mathbf{z}_k = \big[\mathbf{x}_{\boldsymbol{\alpha}}\big]_{\boldsymbol{\alpha} \in \mathbb{Z}_k^d}$ yields the following infinite-dimensional linear dynamical system

$$\dot{\mathbf{z}}_k = \sum_{l=k}^{\infty} \mathbf{A}_{k,l}(t)\, \mathbf{z}_l \tag{2.5}$$

for all $t \geq t_0$ and $k \geq 1$ with initial condition $\mathbf{z}_k(t_0) = \big[\mathbf{x}_{\boldsymbol{\alpha}}^0\big]_{\boldsymbol{\alpha} \in \mathbb{Z}_k^d}$, where

$$\mathbf{A}_{k,l}(t) = \left[\sum_{j=1}^{d} \alpha_j f_{j,\boldsymbol{\beta}-\boldsymbol{\alpha}+\mathbf{e}_j}(t) \right]_{\boldsymbol{\alpha} \in \mathbb{Z}_k^d,\ \boldsymbol{\beta} \in \mathbb{Z}_l^d} \tag{2.6}$$

for all $k, l \geq 1$ are matrices of size $\binom{k+d-1}{d-1} \times \binom{l+d-1}{d-1}$. By defining the infinite-dimensional state vector $\mathbf{z} = \big[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N, \ldots\big]^T$, the set of linear systems (2.5) can be rewritten in compact form

$$\dot{\mathbf{z}} = \mathbf{A}(t)\mathbf{z} \tag{2.7}$$

for $t \geq t_0$ with initial condition $\mathbf{z}(t_0) = \big[\mathbf{x}_{\boldsymbol{\alpha}}^0\big]_{\boldsymbol{\alpha} \in \mathbb{Z}_+^d \setminus \{\mathbf{0}\}}$, where

$$\mathbf{A}(t) = \begin{bmatrix} \mathbf{A}_{1,1}(t) & \mathbf{A}_{1,2}(t) & \cdots & \mathbf{A}_{1,N}(t) & \cdots \\ \mathbf{0} & \mathbf{A}_{2,2}(t) & \cdots & \mathbf{A}_{2,N}(t) & \cdots \\ \vdots & & \ddots & \vdots & \ddots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_{N,N}(t) & \cdots \\ \mathbf{0} & \vdots & & & \ddots \end{bmatrix} \tag{2.8}$$

is a block upper-triangular matrix. The resulting linear system (2.7) is referred to as the Carleman linearization of the nonlinear dynamical system (2.1).

While the original d-dimensional dynamical systems (2.1) is nonlinear, its lifted form (2.5) is an infinite-dimensional linear system whose state matrix $\mathbf{A}$ is an upper-triangular block matrix with special structure and initial condition is of exponential type. On the other hand, the apparent disadvantage of the Carleman linearization is that the resulting state matrix $\mathbf{A}$ is not a bounded operator on $\ell^2(\mathbb{Z}_+^d \setminus \{\mathbf{0}\})$, the Hilbert space of all square-summable sequences on $\mathbb{Z}_+^d \setminus \{\mathbf{0}\}$. Moreover, the initial condition has exponential decay when $\|\mathbf{x}_0\| < 1$ and exponential growth when $\|\mathbf{x}_0\| > 1$, which prevents the direct application of existing theories to analyze the infinite-dimensional linear system on Hilbert spaces $\ell^2(\mathbb{Z}_+^d \setminus \{\mathbf{0}\})$.

A natural question about the original nonlinear system (2.1) and its Carleman linearization (2.5) is how effective the finite section (truncation) of the linearized counterpart is and whether the first component of the solution of the truncated system converges to the solution of the original nonlinear system. Our main contribution shows that if the convergence radius of function $\mathbf{f}(t, \mathbf{x})$ is finite, then the finite section of the Carlemen linearization converges exponentially only when the initial condition is close enough to the equilibrium.

3. CONVERGENCE OF FINITE-SECTIONING OF THE CARLEMAN LINEARIZATION

Denote the bounded norm for $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ by $\|\mathbf{x}\|_\infty = \max_{1 \leq j \leq d} |x_j|$. In this section, we show that the first component of the solution of the finite section approach to the the Carlemen linearization (2.5) converges to the solution of the nonlinear dynamic system (2.1) exponentially when the initial is not too far away from the equilibrium.

The finite section approach to the Carlemen linearization (2.5) can be solved as a linear dynamic system

$$\begin{bmatrix} \dot{\mathbf{y}}_{1,N}(t) \\ \dot{\mathbf{y}}_{2,N}(t) \\ \vdots \\ \dot{\mathbf{y}}_{N,N}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1,1}(t) & \mathbf{A}_{1,2}(t) & \cdots & \mathbf{A}_{1,N}(t) \\ & \mathbf{A}_{2,2}(t) & \cdots & \mathbf{A}_{2,N}(t) \\ & & \ddots & \vdots \\ & & & \mathbf{A}_{N,N}(t) \end{bmatrix} \begin{bmatrix} \mathbf{y}_{1,N}(t) \\ \mathbf{y}_{2,N}(t) \\ \vdots \\ \mathbf{y}_{N,N}(t) \end{bmatrix} \tag{3.9}$$

of dimension $d\binom{N+d}{d} - d$ with initial $\mathbf{y}_{k,N}(t_0) = \left[\mathbf{x}_0^{\boldsymbol{\alpha}}\right]_{\boldsymbol{\alpha}\in\mathbb{Z}_k^d}$ for $1 \le k \le N$.

Assumption 1. The function $\mathbf{f}(t,\mathbf{x})$ in (2.1) is a time-varying analytic function near the origin such that $\mathbf{f}(t,\mathbf{0}) = \mathbf{0}$ for all $t \ge t_0$ and coefficients $\mathbf{f}_{\boldsymbol{\alpha}}(t)$ in its Marclaurin expansion (2.2) satisfy uniform exponential decay property

$$\sup_{t\ge t_0} \sum_{j=1}^{d} \sum_{\boldsymbol{\alpha}\in\mathbb{Z}_k^d} |f_{j,\boldsymbol{\alpha}}(t)| \le D_0 R^{-k} \tag{3.10}$$

for all $k \ge 1$ and some positive constants D_0 and R.

Theorem 2. Suppose that Assumption 1 holds and $\mathbf{x}(t)$ for $t \ge t_0$ is a continuous solution of the nonlinear system (2.1) with initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$ that satisfies

$$0 < \|\mathbf{x}_0\|_\infty < \frac{R}{e}. \tag{3.11}$$

Then, the first component of the solution of the finite section of the Carleman linearization (3.9), i.e., $\mathbf{y}_{1,N}(t)$, converges to $\mathbf{x}(t)$ exponentially as the truncation length N increases for all $t_0 \le t < t_0 + T^*$, i.e., for every $t_0 < t_1 < t_0 + T^*$ there exist a positive constant C such that

$$\sup_{t_0\le t\le t_1} \|\mathbf{y}_{1,N}(t) - \mathbf{x}(t)\|_\infty \le C e^{D_0(t_1-t_0-T^*)N/R} \tag{3.12}$$

for all $N \ge 1$, where

$$T^* = \frac{(e-1)R}{(2e-1)D_0} \ln\left(\frac{R}{e\|\mathbf{x}_0\|_\infty}\right). \tag{3.13}$$

The convergence of the finite section scheme has been studied before by Forets and Pouly (2017), when the right-hand side of the nonlinear system (2.1) is a time-varying polynomial

$$\mathbf{p}_L(t,\mathbf{x}) = \sum_{1\le|\boldsymbol{\alpha}|\le L} \mathbf{p}_{\boldsymbol{\alpha}}(t)\mathbf{x}^{\boldsymbol{\alpha}} \tag{3.14}$$

with degree $L \ge 1$, where $\mathbf{p}_{\boldsymbol{\alpha}}(t) = \left[p_{1,\boldsymbol{\alpha}}(t), \ldots, p_{d,\boldsymbol{\alpha}}(t)\right]^T$. If $L = 1$, it can be verified that the corresponding state matrix $\mathbf{A}(t)$ in the Carlemen linearization (2.8) will be a block diagonal matrix. Hence, the first component $\mathbf{y}_{1,N}(t)$ of the solution of the finite section scheme (3.9) will be equal to the continuous solution $\mathbf{x}(t)$ for all $t \ge t_0$ of the original nonlinear dynamic system (2.1). Now, consider the case that the degree of the polynomial $\mathbf{P}_L$ is at least two, i.e., $L \ge 2$. Define

$$D_0(\mathbf{p}_L, R) = \sup_{1\le k\le L} R^k \sup_{t\ge t_0} \sum_{j=1}^{d} \sum_{\boldsymbol{\alpha}\in\mathbb{Z}_k^d} |p_{j,\boldsymbol{\alpha}}(t)| \tag{3.15}$$

For every $R > 0$, the uniform exponential decay property (3.10) holds for the time-varying polynomial $\mathbf{p}_L(t,\mathbf{x})$ with D_0 replaced by $D_0(\mathbf{p}_L, R)$ and the requirement (3.11) is satisfied for all nonzero initial $\mathbf{x}_0$ when R is chosen appropriately.

Corollary 3. If the right-hand side of system (2.1) is a time-varying polynomial $\mathbf{p}_L(t,\mathbf{x})$ with $L \ge 2$ given by (3.14), then the first component $\mathbf{y}_{1,N}(t)$ of the solution of the truncated system (3.9) will converge to the continuous solution $\mathbf{x}(t)$ of the original nonlinear dynamical system (2.1) for all $t_0 < t < t_0 + T^*(\mathbf{p}_L, e\|\mathbf{x}_0\|_\infty)$, where

$$T^*(\mathbf{p}_L, e\|\mathbf{x}_0\|_\infty) = \sup_{R>s} \frac{(e-1)R}{(2e-1)D_0(\mathbf{p}_L, R)} \ln\left(\frac{R}{e\|\mathbf{x}_0\|_\infty}\right). \tag{3.16}$$

for some $s > 0$.

Let us define quantity

$$a_k := \sup_{t\ge t_0} \sum_{j=1}^{d} \sum_{\boldsymbol{\alpha}\in\mathbb{Z}_k^d} |p_{j,\boldsymbol{\alpha}}(t)|,$$

for all $1 \le k \le L$. One may verify that $D_0(\mathbf{p}_L, R) = a_L R^L$ hold for all $R \ge \max_{1\le k\le L-1}(a_k/a_L)^{1/(L-k)}$. Therefore, when the initial condition satisfies

$$\|\mathbf{x}_0\|_\infty \ge e^{-1} \max_{1\le k\le L-1} \left(\frac{a_k}{a_L}\right)^{1/(L-k)},$$

the maximal achievable time range in (3.16) is

$$T^*(\mathbf{p}_L, e\|\mathbf{x}_0\|_\infty) = \frac{(e-1)(L-1)}{(2e-1)e^{L^2-L} \sup_{t\ge t_0} \sum_{j=1}^{d} \sum_{\boldsymbol{\alpha}\in\mathbb{Z}_L^d} |p_{j,\boldsymbol{\alpha}}(t)|} \|\mathbf{x}_0\|_\infty^{1-L}. \tag{3.17}$$

4. CONCLUSION

For a given time-varying nonlinear system, We proved that under some mild assumptions, the Carelman truncation of the lifted system converges exponentially fast to the original nonlinear system's solution. Explicit error bounds are characterized under the boundedness of the trajectories and initial conditions. Our theoretical results have the potentials to pave the way to prove stability and convergence of Carleman-based methods for optimal control design, e.g., model predictive control of nonlinear systems using the finite truncated system.

REFERENCES

Amini, A., Sun, Q., and Motee, N. (2020a). Approximate optimal control design for a class of nonlinear systems by lifting hamilton-jacobi-bellman equation. In *2020 American Control Conference (ACC)*, 2717–2722. IEEE.

Amini, A., Sun, Q., and Motee, N. (2020b). Quadratization of hamilton-jacobi-bellman equation for near-optimal control of nonlinear systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 731–736. IEEE.

Arnold, V.I., Kozlov, V., and Neishtadt, A. (1988). *Dynamical systems III.* Springer.

Banks, S. (1992). Infinite-dimensional carleman linearization, the lie series and optimal control of non-linear partial differential equations. *International journal of systems science*, 23(5), 663–675.

Bellman, R. and Richardson, J.M. (1963). On some questions arising in the approximate solution of nonlinear differential equations. *Quarterly of Applied Mathematics*, 20(4), 333–339.

Bertsekas, D.P. (2011). *Dynamic programming and optimal control 3rd edition, volume II.* Belmont, MA: Athena Scientific.

Brockett, R. (2014). The early days of geometric nonlinear control. *Automatica*, 50(9), 2203–2224.

Carleman, T. (1932). Application de la théorie des équations intégrales linéaires aux systèmes d'équations différentielles non linéaires. *Acta Mathematica*, 59, 63–87.

Forets, M. and Pouly, A. (2017). *Explicit error bounds for carleman linearization.* arXiv preprint arXiv:1711.02552.

Isidori, A. (2013). *Nonlinear control systems.* Springer Science & Business Media.

Khalil, H.K. (2002). Nonlinear systems third edition. *Patience Hall*, 115.

Koopman, B.O. (1931). Hamiltonian systems and transformation in hilbert space. *Proceedings of the national academy of sciences of the united states of america*, 17(5), 315.

Kowalski, K. and Steeb, W.H. (1991). *Nonlinear dynamical systems and Carleman linearization.* World Scientific.

Steeb, W.H. and Wilhelm, F. (1980). Non-linear autonomous systems of differential equations and carleman linearization procedure. *Journal of Mathematical Analysis and Applications*, 77(2), 601–611.

Svoronos, S., Papageorgiou, D., and Tsiligiannis, C. (1994). Discretization of nonlinear control systems via the carleman linearization. *Chemical engineering science*, 49(19), 3263–3267.

Wiggins, S., Wiggins, S., and Golubitsky, M. (1990). *Introduction to applied nonlinear dynamical systems and chaos*, volume 2. Springer.

Extended Abstracts of the
25th International Symposium on Mathematical Theory of Networks and Systems
Bayreuth, Germany, 12-16 September 2022

Moments, Sums of Squares, and Tropicalization

Grigoriy Blekherman [*] **Felipe Rincón** [**] **Rainer Sinn** [***]
Cynthia Vinzant [****] **Josephine Yu** [†]

[*] *School of Mathematics, Georgia Institute of Technology, Atlanta GA, USA*
[**] *School of Mathematical Sciences, Queen Mary University of London, London, UK*
[***] *Universität Leipzig, Mathematisches Institut, Leipzig, Germany*
[****] *Department of Mathematics, University of Washington, Seattle, WA, USA*
[†] *School of Mathematics, Georgia Institute of Technology, Atlanta GA, USA*

Abstract: The relationship between nonnegative polynomials and sums of squares on semi-algebraic set S is one of the central questions in real algebraic geometry. The (convex) dual side of this story is important in analysis, where it is known as the truncated S-moment problem, and it considers the truncated cones of moments which are dual to nonnegative polynomials, and "pseudo-moments" which are dual to sums of squares. We bring a new tool for understanding of these classical problems: tropicalization. While extensively studied in complex algebraic geometry, tropicalization is rarely applied to semialgebraic sets. We provide explicit combinatorial descriptions of tropicalizations of the moment and pseudo-moment cones, and demonstrate their usefulness in distinguishing between nonnegative polynomials and sums of squares, proving results limiting the power of sums of squares approximations of nonnegative polynomials. We believe that this just scratches the surface of applications of tropicalization in semi-algebraic geometry.

Keywords: Moments, pseudomonents, nonnegative polynomials, sums of squares, inequalities, tropicalization.

1. INTRODUCTION

Understanding nonnegativity of polynomials in terms of sums of squares has been a central challenge in real algebraic geometry dating back to the work of Hilbert. The dual side of this problem is important in analysis and known as the moment problem. We now take a moment to introduce it.

For a semialgebraic set $S \subseteq \mathbb{R}^n$ and a finite subset $A \subset \mathbb{N}^n$, we consider the convex cone $M_A(S)$ of A-moments of measures supported on S. Despite extensive work this cone can be explicitly described in very few situations even when $S = \mathbb{R}^n$ and A corresponds on to all moments of degree at most $2d$ Curto and Fialkow (1996, 1991); di Dio and Schmüdgen (2018); Schmüdgen (2017). An important tool for understanding $M_A(S)$ comes from *Positivstellensätze* in real algebraic geometry: theorems on representing the dual cone of polynomials with support in A which are nonnegative on S via sums of squares Schmüdgen (1991); Putinar (1993). We denote the cone of linear functionals dual to the cone of "obviously nonnegative" polynomials generated by sums of squares by $\Sigma(S)_A^\vee$ and call such functionals "pseudo-moments". Tropicalization of the cones of moments and pseudo-moments gives us "combinatorial shadows" of these sets. Our explicit descriptions of these shadows lead to interesting combinatorial questions, some of which have been considered in the context of SONC polynomials Reznick (1989); Iliman and de Wolff (2016); Katthän et al. (2021).

Another way of understanding our results is through binomial inequalities in moments and pseudo-moments of measures supported on S. When the semialgebraic set S is closed under Hadamard multiplication, the tropicalization $\operatorname{trop} M_A(S)$ of the moment cone is a rational polyhedral cone. Its dual cone $(\operatorname{trop} M_A(S))^\vee$ encodes all of the binomial inequalities in A-moments. Similarly, binomial moment inequalities that can proved via sums of squares correspond to another rational polyhedral cone, which may depend on a degree bound for the sums of squares construction. While polynomial inequalities valid on $M_A(S)$ are difficult to characterize, we explicitly describe all binomial inequalities in moments and pseudo-moments by finding the extreme rays of the corresponding rational polyhedral cones. The use of tropicalizations to analyze the power of sums of squares method was first introduced in Blekherman et al. (2020) for analyzing graph density inequalities, and further developed in Blekherman and Raymond (2021). We take inspiration from some of their results and techniques, for instance the use of the

⋆ GB is partially supported by US National Science Foundation grant DMS-1901950. JY is partially supported by US National Science Foundation grant #1855726.

Hadamard property to ensure that the tropicalization is a convex cone. However, to the best of our knowledge, this is the first instance where tropicalization is used to study the relationship between the moment and pseudo-moment cones.

We start with a pair of examples which illustrate our setup and results

Example 1.1. (Motzkin Configuration on Orthant). Let $S = \mathbb{R}^2_{\geq 0}$ be the nonnegative orthant and let A be the *Motzkin configuration*: $A = \{(0,0),(1,2),(2,1),(1,1)\}$, which gives us the exponents of moments we are recording:

$$m_{00} = \int_S 1\, d\mu,\ m_{12} = \int_S xy^2\, d\mu,$$
$$m_{21} = \int_S x^2 y\, d\mu,\ m_{11} = \int_S xy\, d\mu.$$

There is only one binomial inequality satisfied by A-moments of measures supported on S:

$$m_{00}m_{12}m_{21} \geq m_{11}^3. \tag{1}$$

If we regard moments as functions on A, then we see that moments are nonnegative *log-convex functions* on A, and in fact inequalities coming from log-convexity are the only possibly binomial inequalities in A-moments for measures supported on the nonnegative orthant $\mathbb{R}^n_{\geq 0}$ (see thm:genmom).

We now consider A-pseudo-moments of measures supported on $\mathbb{R}^n_{\geq 0}$. Pseudo-moments are defined as linear functionals that are nonnegative on "obviously" nonnegative polynomials coming from sums of squares (see sec:SOS). We show in tropicalizationpseudononneg that A-pseudo-moments of measures supported on $\mathbb{R}^n_{\geq 0}$ satisfy *log-midpoint-convexity inequalities*:

$$m_\alpha m_\beta \geq m^2_{\left(\frac{\alpha+\beta}{2}\right)}, \tag{2}$$

with $\alpha, \beta, \frac{\alpha+\beta}{2} \in A$. Moreover these inequalities generate all possible binomial inequalities valid on A-pseudomoments. Since the Motzkin configuration contains no midpoints, we see that there are no binomial inequalities valid on A-pseudomoments. ◇

Remark 1.2. The combinatorial notions of convex and midpoint-convex functions on A are quite similar to what has been developed for analyzing certain sparse globally nonnegative polynomials and sums of squares arising from the arithmetic mean-geometric mean inequality. Such polynomials were originally called AGI-forms by Reznick in Reznick (1989) and were later called Sum of Nonnegative Circuit Polynomials (SONC) in Iliman and de Wolff (2016). The only difference is that for analyzing global nonnegativity, it makes a difference whether points in A have all even coordinates or not, and for instance midpoints convexity has to hold only between even points in A. As we will see in thm:whole and tropicalizingpseudomoments this is precisely what happens for us as well when analyzing measures supported on all of $\mathbb{R}^n$. ◇

Example 1.3. (Motzkin Configuration on Square.). Let $S = [0,1]^2 \subset \mathbb{R}^2$ be the unit square given by inequalities $0 \leq x \leq 1$, $0 \leq y \leq 1$. Let $A \subset \mathbb{N}^2$ again be the Motzkin configuration. In addition to the log-convexity inequality (1) the following binomial moment inequalities are naturally valid on the unit square, since all variables lie between 0 and 1:

$$m_{11} \geq m_{12},\ m_{11} \geq m_{21}\ m_{00} \geq m_{11}.$$

Any binomial inequality in A-moments of measures supported on S can be obtained from the above inequalities and (1) via exponentiation and multiplication (See Example 1.3).

As we increase the degree d, sums of squares provide increasingly better approximations to polynomials supported on A that are nonnegative on S, and thus can, in principle, be used to provide increasingly sharper binomial inequalities for pseudo-moments. If we regard pseudo-moments as functions on A then increasing the degree allows us to use moments that lie outside of A. For instance we can show that $m_{00}m_{12} \geq 2m_{11}$ by combining the inequality $m_{12} \geq m_{22}$ with the log-midpoint-convexity inequality (2): $m_{00}m_{22} \geq m_{11}^2$.

We show that the binomial A-pseudo-moment inequalities *stabilize*, and only the following binomial inequalities can be learned via sums of squares (regardless of the degree d):

$$m_{11} \geq m_{12},\ \ m_{11} \geq m_{21},\ \ m_{00}m_{12} \geq m_{11}^2,$$
$$\text{and}\ \ m_{00}m_{21} \geq m_{11}^2.$$

Therefore for any degree d sums of squares cannot prove the moment inequality $m_{00}m_{12}m_{21} \geq m_{11}^3$, and moreover, sums of squares remain *quantifiably far away* from certifying this inequality. ◇

Remark 1.4. Since the unit square is compact, it follows that from Schmüdgen's Positivstellensatz Schmüdgen (1991) that any polynomial f strictly positive on the unit square has a sum of squares certificate. Therefore, as degree increases sums of squares provide an increasingly better approximation to all nonnegative polynomials supported on A. However, as we have seen, tropicalizations stabilize, and higher degree sums of squares do not have larger tropicalizations. This is due to the fact that $\mathrm{trop}(S)$ only depends on the neighborhood of zero and the "neighborhood of infinity" contained in S. We give a simple example of this phenomenon below: Let S be the planar triangle with vertices $(0,0)$, $(1,0)$ and $(1,1)$ and let S_ε be the quadrilateral with vertices $(0,0)$, $(1,0)$, $(1,1)$ and $(0,\varepsilon)$. The we have $S_\varepsilon \to S$ as $\varepsilon \to 0$, however $\mathrm{trop}(S_\varepsilon)$ is the entire plane for all $\varepsilon > 0$. ◇

Remark 1.5. The unit square is special in that all nonnegative polynomials have a sum of squares certificate. Example 1.3 also shows that even though every nonnegative polynomial is a sum of squares, there does not exist a degree bound for the certificate even for just the A-supported polynomials (Marshall, 2008, 9.4.6 Example (1)). ◇

1.1 Main Results in Detail:

We say that a subset S of $\mathbb{R}^n_{\geq 0}$ has the Hadamard property if S is closed under coordinatewise (Hadamard multiplication). Our main results are about the tropicalizations

of moment cones and pseudo-moment cones for semialgebraic sets with the Hadamard property. Concretely, we focus on nonnegative orthants, hypercubes, and toric cubes to discuss our general results. Throughout, we fix a finite set $A \subset \mathbb{Z}^n_{\geq 0}$ of exponents and consider the A-moments, that is to say $\int_S x^a$ for $a \in A$ (also known as truncated moment sequences).

We think of elements of the tropicalization of the moment cone (resp. pseudo-moment cone) as functions $h\colon A \to \mathbb{R}$ and describe the tropicalization mainly in terms of discrete convexity properties of these functions. For the moment cone, we have a general description of the tropicalization of $M_A(S)$ for any subset of the nonnegative orthant with the Hadamard property:

Theorem 1.6. Let $S \subset \mathbb{R}^n_{\geq 0}$ be a semialgebraic set with the Hadamard property such that the intersection of S with the positive orthant is dense in S. The tropicalization of the A-moment cone $M_A(S)$ is the rational polyhedral cone of functions $h\colon A \to \mathbb{R}$ satisfying the following linear inequalities:

(1) (Convexity:) $\sum_{i=1}^r \lambda_i h(a_i) \geq h(b)$ for all $a_1, \ldots, a_r, b \in A$, $\lambda_i \geq 0$, $\sum_{i=1}^r \lambda_i = 1$;
(2) (Nonincreasing:) $h(a) \geq h(b)$ whenever $a - b \in \operatorname{trop}(S)^\vee$.

The first type of inequality is the naive form of discrete convexity that arises in this context. The second type of inequality is where the set S enters: The tropicalization of S is a rational polyhedral cone and $\operatorname{trop}(S)^\vee$ is its dual cone, which defines a partial order of $\mathbb{R}^A$ – and the second inequality says that the functions in the tropicalization are order preserving in this sense. Below, we combine these two types by writing the inequality description as $\sum \lambda_i h(a_i) \geq h(b)$ whenever $\sum \lambda_i a_i - b \in \operatorname{trop}(S)^\vee$. In case $S = \mathbb{R}^n_{\geq 0}$, we have $\operatorname{trop}(S)^\vee = \{0\}$ so that the tropicalizations of the A-moment cones do not include inequalities of type (2). For $S = [0,1]^n$, we get $\operatorname{trop}(S)^\vee = \mathbb{R}^n_{\leq 0}$ and the inequalities of type (2) say that the functions $h \in \operatorname{trop}(M_A(S))$ are non-increasing in coordinate directions.

We can also think of this result as an elegant description of all binomial inequalities valid on the moment (by exponentiation). With the analogous result for pseudo-moment cones, we will see that these inequalities suffice in distinguishing moments from pseudo-moments in many important cases. Moreover, there is a rich combinatorial interplay between geometry of the moment configuration A and geometry and algebraic description of S.

We now move on to pseudo-moment cones, which are the dual cones to truncated preorderings or quadratic modules. We describe in detail how we truncate (in a total degree version) at the beginning of sec:tsos. For pseudo-moment cones, we focus on the case that the semialgebraic set S has an inequality description in terms of pure binomial inequalities.

Theorem 1.7. Let $S \subset \mathbb{R}^n_{\geq 0}$ be a semi-algebraic set defined by pure binomial inequalities $g_i = x^{a_i} - x^{b_i}$ such that $S \subset \overline{S \cap \mathbb{R}^n_{>0}}$. Assume that the exponent vectors $a_i - b_i$ of the binomials defining S generate the semigroup $N = \operatorname{trop}(S)^\vee \cap \mathbb{Z}^n$. For all sufficiently large d the tropicalization of $\mathrm{QM}_d(g_i)^\vee$ is the rational polyhedral cone $F(S)_d$ given by the following inequalities:

(1) (Midpoint convexity:) $h(a_1) + h(a_2) \geq 2h(b)$ for all a_1, a_2, b such that $|a_i| \leq d$, $|b| \leq d$ and $a_1 + a_2 = 2b$;
(2) (Nonincreasing:) $h(a) \geq h(b)$ whenever $a - b \in \operatorname{trop}(S)^\vee$.

The inequalities in A-pseudo-moments provable by sums of squares of degree at most d are dual to the coordinate projection of $F(S)_d$ onto the coordinates of A.

In the case of pseudo-moments, we need the additional assumption on the inequality description of S that the exponent vectors of the inequalities generate the semigroup of lattice points in the convex cone $\operatorname{trop}(S)^\vee$ to give the same inequalities of type (2) as in the case of moment cones. This is an assumption that, from a purely theoretical point of view, can be made without loss of generality by adding valid and redundant inequalities, if necessary. Without this assumption, we only get some inequalities of type (2), namely those corresponding to the lattice points in $\operatorname{trop}(S)^\vee$ that also lie in the semigroup generated by the exponent vectors.

Our most intriguing observation is that tropicalizations of pseudomoment cones *stabilize* as the degree bound d grows. This means that for sufficiently large d the tropicalizations of pseudomoment cones remain the same, even though pseudomoments themselves provide a convergent approximation to the moment cone. This phenomenon was already observed in Example 1.3. We provide an explicit description of when stabilization occurs for the hypercube $[0,1]^n$ and provide examples of stabilization and a general theorem.

REFERENCES

Blekherman, G. and Raymond, A. (2021). A path forward: Tropicalization in extremal combinatorics. *arXiv preprint arXiv:2108.06377*.

Blekherman, G., Raymond, A., Singh, M., and Thomas, R.R. (2020). Tropicalization of graph profiles. *arXiv preprint arXiv:2004.05207, to appear in Transactions of the American Mathematical Society*.

Curto, R.E. and Fialkow, L.A. (1991). Recursiveness, positivity, and truncated moment problems. *Houston J. Math.*, 17(4), 603–635.

Curto, R.E. and Fialkow, L.A. (1996). Solution of the truncated complex moment problem for flat data. *Mem. Amer. Math. Soc.*, 119(568), x+52. doi:10.1090/memo/0568. URL `https://doi.org/10.1090/memo/0568`.

di Dio, P.J. and Schmüdgen, K. (2018). The multidimensional truncated moment problem: atoms, determinacy, and core variety. *J. Funct. Anal.*, 274(11), 3124–3148. doi:10.1016/j.jfa.2017.11.013. URL `https://doi.org/10.1016/j.jfa.2017.11.013`.

Iliman, S. and de Wolff, T. (2016). Amoebas, nonnegative polynomials and sums of squares supported on circuits. *Res. Math. Sci.*, 3, Paper No. 9, 35.

Katthän, L., Naumann, H., and Theobald, T. (2021). A unified framework of sage and sonc polynomials and its duality theory. *Mathematics of Computation*, 90(329), 1297–1322.

Marshall, M. (2008). *Positive polynomials and sums of squares*, volume 146 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI. doi:10.1090/surv/146. URL https://doi.org/10.1090/surv/146.

Putinar, M. (1993). Positive polynomials on compact semi-algebraic sets. *Indiana Univ. Math. J.*, 42(3), 969–984. doi:10.1512/iumj.1993.42.42045. URL https://doi.org/10.1512/iumj.1993.42.42045.

Reznick, B. (1989). Forms derived from the arithmetic-geometric inequality. *Mathematische Annalen*, 283(3), 431–464.

Schmüdgen, K. (1991). The K-moment problem for compact semi-algebraic sets. *Math. Ann.*, 289(2), 203–206. doi:10.1007/BF01446568. URL https://doi.org/10.1007/BF01446568.

Schmüdgen, K. (2017). *The moment problem*, volume 277 of *Graduate Texts in Mathematics*. Springer, Cham.